Optimal Forgery and Suppression of Ratings for Privacy Enhancement in Recommendation Systems

Javier Parra-Arnau, David Rebollo-Monedero and Jordi Forné

Abstract

Recommendation systems are information-filtering systems that tailor information to users on the basis of knowledge about their preferences. The ability of these systems to profile users is what enables such intelligent functionality, but at the same time, it is the source of serious privacy concerns. In this paper we investigate a privacy-enhancing technology that aims at hindering an attacker in its efforts to accurately profile users based on the items they rate. Our approach capitalizes on the combination of two perturbative mechanisms—the forgery and the suppression of ratings. While this technique enhances user privacy to a certain extent, it inevitably comes at the cost of a loss in data utility, namely a degradation of the recommendation's accuracy. In short, it poses a trade-off between privacy and utility.

The theoretical analysis of said trade-off is the object of this work. We measure privacy as the Kullback-Leibler divergence between the user's and the population's item distributions, and quantify utility as the proportion of ratings users consent to forge and eliminate. Equipped with these quantitative measures, we find a closed-form solution to the problem of optimal forgery and suppression of ratings, and characterize the optimal trade-off surface among privacy, forgery rate and suppression rate. Experimental results on a popular recommendation system show how our approach may contribute to privacy enhancement.

Index Terms

Information privacy, Kullback-Leibler divergence, user profiling, privacy-enhancing technologies, data perturbation, recommendation systems.

I. Introduction

From the advent of the Internet and the World Wide Web, the amount of information available to users has grown exponentially. As a result, the ability to find information relevant for their interests has become a central issue in recent years. In this context of information overload, *recommendation systems* arise to provide information tailored to users on the basis of knowledge about their preferences [2]. In essence, a recommendation system may be regarded as a type of information-filtering system that suggests information items users may be interested in. Examples of such systems include recommending music at *Last.fm* and *Pandora Radio*, movies by *MovieLens* and *Netflix*, videos at *YouTube*, news at *Digg* and *Google News*, and books and other products at *Amazon*.

Most of these systems capitalize on the creation of *profiles* that represent interests and preferences of users. Such profiles are the result of the collection and analysis of the data that users communicate to those systems. A distinction is frequently made between *explicit* and *implicit* forms of data collection. The most popular form of explicit data collection is that users communicate their preferences by rating items. This is the case of many of the applications mentioned above, where users assign *ratings* to songs, movies or news they have already listened, watched or read. Other strategies to capture users' interests include asking them to sort a number of items by order of predilection, or suggesting that they mark the items they like. On the other hand, recommendation systems may collect data from users without requiring them to explicitly convey their preferences [3]. These practices comprise observing the items clicked by users in an online store, analyzing the time it takes users to examine an item, or simply keeping a record of the purchased items.

The prolonged collection of these personal data allows the system to extract an accurate snapshot of user interests, i.e., their profiles. With this invaluable source of information, the recommendation system applies some technique [4] to generate a prediction of users' interests for those items they have not yet considered. For example, *Movielens* and *Digg* use collaborative-filtering techniques to predict the rating that a user would give to a movie and to create a personalized list of recommended news, respectively. In a nutshell, the ability of profiling users based on such personal information is precisely what enables the intelligent functionality of those systems.

Despite the many advantages recommendation systems are bringing to users, the information collected, processed and stored by these systems prompts serious privacy concerns. One of the main privacy risks perceived by users is that of a computer "figuring things out" about them [5]. Many users are worried about the idea that their profiles may reveal sensitive information such as health-related issues, political preferences, salary or religion. Such privacy risk is exacerbated especially when these profiles are combined across several information services or enriched with data from social networks. An illustrative example

Some parts of this paper (a reduced version of Secs. I and II) were presented at the International Workshop on Data Privacy Management, Leuven, Belgium, Sep. 2011 [1]. The formulation of the trade-off between privacy and utility (Sec. III), the theoretical analysis (Sec. IV), the experiments (Sec. V) and the conclusions (Sec. VI) are all new work.

The authors are with the Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain (e-mail: javier.parra@entel.upc.edu; david.rebollo@entel.upc.edu; jforne@entel.upc.edu).

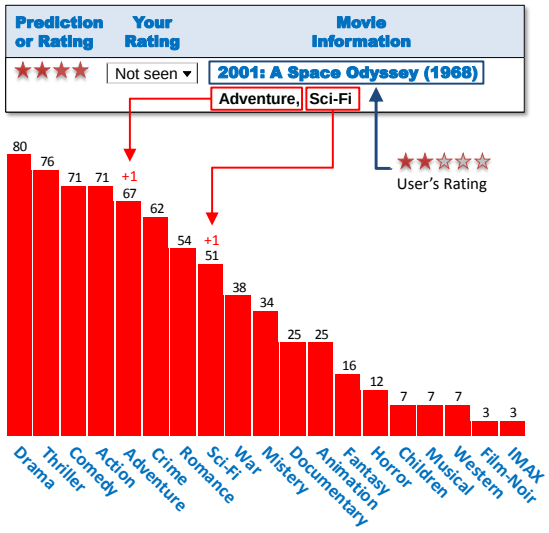


Fig. 1: The profile of a user is modeled in *Movielens* as a histogram of absolute frequencies of ratings within a set of movie genres (bottom). Based on this profile, the recommender predicts the rating that the user would probably give to a movie (top). After having watched the movie, the user rates it and their profile is updated.

is [6], which demonstrates that it is possible to unveil sensitive information about a person from their movie rating history by cross-referencing data from other sources. The authors analyzed the *Netflix Prize* data set [7], which contained anonymous movie ratings of around half a million users of *Netflix*, and were able to uncover the identity, political leaning and even sexual orientation of some of those users, by simply correlating their ratings with reviews they posted on the popular movie Web site *IMDb*. Apart from the risk of cross-referencing, users are also concerned that the system's predictions may be totally erroneous and be later used to defame them. This latter situation is examined in [8], where the accuracy of the predictions provided by *TiVo* digital video recorder and *Amazon* is questioned. Lastly, other privacy risks embrace unsolicited marketing, information leaked to other users of the same computer, court subpoenas, and government surveillance [5].

As a result of all this, it is not surprising that some users are reticent to reveal their interests. In fact, [9] reports that the 24% of Internet users surveyed provided false information in order to avoid giving private information to a Web site. Alternatively, another study [10] finds that 95% of the respondents refused, at some point, to provide personal information when requested by a Web site. In closing, these studies seem to indicate that submitting false information and refusing to give private information are strategies accepted by users concerned with their privacy.

A. Contribution and Plan of this Paper

In this paper we approach the problem of protecting user privacy in those recommendation systems that profile users on the basis of the items they rate. Given the willingness of users to provide fake information and elude disclosing private data, we investigate a privacy-enhancing technology (PET) that combines these two forms of data perturbation, namely the forgery and the suppression of ratings. Concordantly, in our scenario users rate those items they have an opinion on. However, in order to avoid being accurately profiled by the recommender or, in general, by any privacy attacker capable of collecting this information, users may wish to refrain from rating some of those items and/or rate items that do not reflect their actual preferences. Our approach thus protects user privacy to a certain degree, without having to trust the recommendation system or the network operator, but at the cost a loss in utility, a degradation of the quality of the recommendation. In other words, our PET poses a trade-off between privacy and utility.

The theoretical analysis of the trade-off between these two contrasting aspects is the object of this work. We tackle the issue in a systematic fashion, drawing upon the methodology of multiobjective optimization. Before proceeding, though, we adopt a quantifiable measure of user privacy—the Kullback-Leibler (KL) divergence between the probability distribution of the user's items and the population's distribution, a criterion that we introduced in previous work [11] and justified and interpreted in [12], [13] by leveraging on the rationale behind entropy-maximization methods. Equipped with a measure of both privacy and utility, we formulate an optimization problem modeling the trade-off between privacy on the one hand, and on the other forgery rate and suppression rate as utility metrics. Our extensive theoretical analysis finds a closed-form solution to the problem of optimal forgery and suppression of ratings, and characterizes the optimal trade-off between the aspects of privacy and utility.

In addition, we provide an empirical evaluation of our data-perturbative approach. Specifically, we apply the forgery and the suppression of ratings in the popular movie recommendation system *Movielens*, and show how these two strategies may preserve the privacy of its users.

Sec. II reviews several data-perturbative approaches aimed at enhancing user privacy in the context of recommender systems. Sec. III introduces our privacy-enhancing technology, proposes a quantitative measure of the privacy of user profiles, and formulates the trade-off between privacy and utility. Sec. IV presents a theoretical analysis of the optimization problem

characterizing the privacy-forgery-suppression trade-off. In this same section we also provide a numerical example that illustrates our formulation and theoretical results. Sec. V evaluates our privacy-protecting mechanism in a real recommendation system. Finally, conclusions are drawn in Sec. VI.

II. State of the Art

Numerous approaches have been proposed to protect user privacy in the context of recommendation systems. These approaches fundamentally suggest either perturbing the information provided by users or using cryptographic techniques.

In the case of perturbative methods for recommendation systems, [14] proposes that users add random values to their ratings and then submit these perturbed ratings to the recommender. After receiving these ratings, the system executes an algorithm and sends the users some information that allows them to compute the prediction. When the number of participating users is sufficiently large, the authors find that user privacy is protected to a certain extent and the system reaches a decent level of accuracy. However, even though a user disguises all their ratings, it is evident that the items themselves may uncover sensitive information. Simply put, the mere fact of showing interest in a certain item may be more revealing than the rating assigned to that item. For instance, a user rating a book called "How to Overcome Depression" indicates a clear interest in depression, regardless of the score assigned to this book. Apart from this critique, other works [15], [16] stress that the use of *randomized* data distortion techniques might not be able to preserve privacy.

In line with this work, [17] applies the same data-perturbative technique to collaborative-filtering algorithms based on singular-value decomposition. Specifically, the authors focus on the impact that their technique has on privacy. For this purpose, they use the privacy metric proposed by [18], which is essentially equivalent to differential entropy, and conduct some experiments with data sets from *Movielens* and *Jester*. The results show the trade-off curve between accuracy in recommendations and privacy. In particular, they measure accuracy as the mean absolute error between the predicted values from the original ratings and the predictions obtained from the perturbed ratings.

At this point, we would like to remark that the use of perturbative techniques is by no means new in other scenarios such as private information retrieval and the semantic Web. In the former scenario, users send general-purpose queries to an information service provider. A perturbative approach to protect user profiles in this context consists in combining genuine with false queries. Precisely, [11] proposes a *nonrandomized* method for query forgery and investigates the trade-off between privacy and the additional traffic overhead. In the semantic Web scenario, users annotate resources with the purpose of classifying them. In this application domain, the perturbation of user profiles for privacy preservation may be carried out by dropping certain annotations or *tags*. An example of this kind of perturbation may be found in [19]–[21], where the authors propose the elimination of tags as a privacy-enhancing strategy.

Regarding the use of cryptographic techniques, [22], [23] propose a method that enables a community of users to calculate a public aggregate of their profiles without revealing them on an individual basis. In particular, the authors use a homomorphic encryption scheme and a peer-to-peer communication protocol for the recommender to perform this calculation. Once the aggregated profile is computed, the system sends it to users, who finally use local computation to obtain personalized recommendations. This proposal prevents the system or any external attacker from ascertaining the individual user profiles. However, its main handicap is assuming that an acceptable number of users is online and willing to participate in the protocol. In line with this, [24] uses a variant of Pailliers' homomorphic cryptosystem which improves the efficiency in the communication protocol. Another solution [25] presents an algorithm aimed at providing more efficiency by using the scalar product protocol.

III. Privacy Protection via Forgery and Suppression of Ratings

In this section, first we present the forgery and the suppression of ratings as a privacy-enhancing technology. The description of our approach is prefaced by a brief introduction of the concepts of soft privacy and hard privacy. Secondly, we propose a model of user profile and set forth our assumptions about the adversary capabilities. Finally, we provide a quantitative measure of both privacy and utility, and present a formulation of the trade-off between these two contrasting aspects.

A. Soft Privacy vs. Hard Privacy

The privacy research literature [26] recognizes the distinction between the concepts of *soft privacy* and *hard privacy*. A privacy-enhancing mechanism providing *soft privacy* assumes that users entrust their private data to an entity, which is thereafter responsible for the protection of their data. In the literature, numerous attempts to protect privacy have followed the traditional method of anonymous communications [27]–[30], which is fundamentally based on the suppositions of soft privacy. Unfortunately, anonymous-communication systems are not completely effective [31]–[34], they normally come at the cost of infrastructure, and assume that users are willing to trust other parties.

Our privacy-protecting technique, per contra, leverages on the principle of *hard privacy*, which assumes that users mistrust communicating entities and therefore strive to reveal as little private information as possible. In the motivating scenario of this work, hard privacy means that users need not trust an external entity such as the recommender or the network operator. Consequently, because users just trust themselves, it is their own responsibility to protect their privacy. In this state of affairs, the forgery and the suppression of ratings appear as a technique that may hinder privacy attackers in their efforts to accurately

profile users on the basis of the items they rate. Specifically, when users are adhered to this technique, they have the possibility to submit ratings to items that do not reflect their genuine preferences, and/or refrain from rating some items of their interest—this is what we refer to as the *forgery* and the *suppression* of ratings, respectively.

B. User Profile and Adversary Model

In the scenario of recommendation systems, users rate items of a very different nature, e.g., music, pictures, videos or news, according to their personal preferences. The information conveyed allows those systems to extract a profile of interests or *user profile*, which turns to be essential in the provision of personalized recommendations.

We mentioned in Sec. I that *Movielens* represents user profiles by using some kind of histogram. Other systems such as *Jinni* and *Last.fm* show this information by means of a tag cloud, which in essence may be regarded as another kind of histogram. In this same spirit, recent privacy-protecting approaches in the scenario of recommendation systems also propose using histograms of absolute frequencies for modeling user profiles [35], [36].

According to these examples and inspired by other works in the field [1], [11], [19]–[21], [37], we model the *items* rated by users as random variables (r.v.'s) taking on values in a common finite alphabet of categories, namely the set $\{1, \ldots, n\}$ for some integer $n \geqslant 2$. Concordantly, we model the profile of a user as a probability mass function (PMF) $q = (q_1, \ldots, q_n)$, that is, a histogram of relative frequencies of items within a predefined set of categories of interest.

We would like to emphasize that, under this model, user profiles do not capture the particular scores given to items, but what we consider to be more sensitive: the categories these items belong to. This is exactly the case of *Movielens* and numerous content-based recommendation systems. Fig. 1 provides an example that illustrates how user profiles are constructed in *Movielens*. In this particular example, a user assigns two stars to a movie, meaning that they consider it to be "fairly bad". However, the recommender updates their profile based only on the categories this movie belongs to.

According to this model, a privacy attacker supposedly observes a perturbed version of this profile, resulting from the forgery and the suppression of certain ratings, and is unaware or ignores the fact that the observed user profile, also in the form of a histogram, does not reflect the actual profile of interests of the user in question. In principle, our passive attacker could be the recommender itself or the network operator. However, the set of potential attackers is not restricted merely to these two entities. Since ratings are often publicly available to other users of the recommendation system, any other attacker able to crawl through this information is taken into consideration in our adversary model.

When users adhere to the forgery and the suppression of ratings, they specify a *forgery rate* $\rho \in [0, \infty)$ and a *suppression rate* $\sigma \in [0, 1)$. The former is the ratio of forged ratings to total genuine ratings that a user consents to submit. The latter ratio is the fraction of genuine ratings that the user agrees to eliminate [(a)]. Note that, in our approach, the number of false ratings submitted by the user can exceed the number of genuine ratings, that is, ρ can be greater than 1. Nevertheless, the number of suppressed ratings is always lower than the number of genuine ratings.

By forging and suppressing ratings, the *actual* profile of interests q is then perceived from the outside as the *apparent* PMF $t = \frac{q+r-s}{1+\rho-\sigma}$, according to a *forgery strategy* $r = (r_1, \ldots, r_n)$ and a *suppression strategy* $s = (s_1, \ldots, s_n)$. Such strategies represent the proportion of ratings that the user should forge and eliminate in each of the n categories. Naturally, these strategies must satisfy, on the one hand, that $r_i \geqslant 0$, $s_i \geqslant 0$ and $q_i + r_i - s_i \geqslant 0$ for $i = 1, \ldots, n$, and on the other, that $\sum_{i=1}^{n} r_i = \rho$ and $\sum_{i=1}^{n} s_i = \sigma$. In conclusion, the apparent profile is the result of the addition and the substraction of certain items to/from the actual profile, and the posterior normalization by $\frac{1}{1+\rho-\sigma}$ so that $\sum_{i=1}^{n} t_i = 1$.

C. Measuring the Privacy of User Profiles

Inspired by the privacy measures proposed in [11]–[13], [19], [38], and according to the model of user profile assumed in Sec. III-B, we define *initial privacy risk* as the KL divergence [39] between the user's genuine profile and the population's distribution, that is,

$$\mathcal{R}_0 = \mathrm{D}(q \,\|\, p).$$

Similarly, we define *(final) privacy risk* $\mathcal{R}$ as the KL divergence between the user's apparent profile and the population's distribution,

$$\mathcal{R} = \mathrm{D}(t \,\|\, p) = \mathrm{D}\left(\frac{q+r-s}{1+\rho-\sigma} \,\middle\|\, p\right).$$

An intuitive justification of our privacy metric stems from the observation that, whenever the user's apparent item distribution diverges too much from the population's, a privacy attacker will have actually gained some information about the user, in contrast to the statistics of the general population.

A richer argument may be found in [12], [13], where we establish some riveting connections between Jaynes' rationale on entropy-maximization methods and the use of entropies and divergences as measures of privacy. The leading idea is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a PMF

[(a)]The description of an architecture implementing this data-perturbative approach may be found in [1].

in a stochastic system and its Shannon's entropy. Loosely speaking and in our context, the higher the entropy of a profile, the more likely it is, the more users behave similarly. This is in absence of a probability distribution model for the PMFs, viewed abstractly as r.v.'s themselves. Under this interpretation, Shannon's entropy is a measure of anonymity, *not* in the sense that the user's identity remains unknown, but only in the sense that higher likelihood of an apparent profile, believed by an external observer to be the actual profile, makes that profile more common, helping the user go unnoticed, less interesting to an attacker assumed to strive to target peculiar users.

If an aggregated histogram of the population were available as a reference profile, as we assume in this work, the extension of Jaynes' argument to relative entropy also gives an acceptable measure of privacy (or anonymity). Recall [39] that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. Conceptually, a lower KL divergence hides discrepancies with respect to a reference profile, say the population's, and there also exists a monotonic relationship between the likelihood of a distribution and its divergence with respect to the reference distribution of choice, which enables us to regard KL divergence as a measure of anonymity in a sense entirely analogous to the above mentioned.

D. Formulation of the Trade-Off among Privacy, Forgery and Suppression

Our data-perturbative mechanism allows users to enhance their privacy to a certain extent, since the resulting profile, as observed from the outside, no longer captures their actual interests. The price to be paid, however, is a loss in data utility, in particular in the accuracy of the recommender's predictions.

For the sake of tractability, in this work we consider as utility metrics the forgery rate and the suppression rate. This consideration enables us to formulate the problem of choosing a forgery strategy and a suppression strategy as a multiobjective optimization problem that takes into account privacy, forgery rate and suppression rate. Specifically, under the assumption that the population of users is large enough to neglect the impact of the choice of r and s on p, we define the *privacy-forgery-suppression* function

$$\mathcal{R}(\rho,\sigma) = \min_{\substack{r,s \\ r_i \geqslant 0,\, s_i \geqslant 0, \\ q_i + r_i - s_i \geqslant 0, \\ \sum r_i = \rho,\, \sum s_i = \sigma}} \mathrm{D}\left(\frac{q + r - s}{1 + \rho - \sigma} \,\middle\|\, p \right), \tag{1}$$

which characterizes the optimal trade-off among privacy, forgery rate and suppression rate.

Conceptually, the result of this optimization are two strategies r and s that contain information about which ratings should be forged and which ones should be suppressed, in order to achieve the minimum privacy risk. More precisely, the component r_i is the percentage of items that the user should forge in the category i. The component s_i is defined analogously for suppression.

IV. Optimal Forgery and Suppression of Ratings

This section is entirely devoted to the theoretical analysis of the privacy-forgery-suppression function (1) defined in Sec. III-D. In our attempt to characterize the trade-off among privacy risk, forgery rate and suppression rate, we shall present a closed-form solution to the optimization problem inherent in the definition of this function. Afterwards, we shall analyze some fundamental properties of said trade-off. For the sake of brevity, our theoretical analysis only contemplates the case when all given probabilities are strictly positive:

$$q_i, p_i > 0 \text{ for all } i = 1, \ldots, n. \tag{2}$$

Additionally, we suppose without loss of generality that

$$\frac{q_1}{p_1} \leqslant \cdots \leqslant \frac{q_n}{p_n}. \tag{3}$$

Before diving into the mathematical analysis, it is immediate from the definition of the privacy-forgery-suppression function that its initial value is $\mathcal{R}(0,0) = \mathrm{D}(q \,\|\, p)$. The characterization of the optimal trade-off surface modeled by $\mathcal{R}(\rho,\sigma)$ at any other values of ρ and σ is the focus of this section.

A. Closed-Form Solution

Our first theorem, Theorem 3, will present a closed-form solution to the minimization problem involved in the definition of function (1). The solution will be derived from Lemma 1, which addresses a resource allocation problem. This a theoretical problem encountered in many fields, from load distribution and production planning to communication networks, computer scheduling and portfolio selection [40]. Although this lemma provides a parametric-form solution, we shall be able to proceed towards an explicit closed-form solution, albeit piecewise.

Lemma 1 (Resource Allocation): For all $k = 1, \ldots, n$, let f_k be a real-valued function on $\{(x_k, y_k) \in \mathbb{R}^2 \colon \kappa_k + x_k - y_k \geqslant 0\}$, twice differentiable in the interior of its domain. Assume that $\frac{\partial f_k}{\partial x_k} = -\frac{\partial f_k}{\partial y_k}$, that $\frac{\partial^2 f_k}{\partial x_k^2} = \frac{\partial^2 f_k}{\partial y_k^2} > 0$ and that the Hessian $H(f_k)$ is positive semidefinite. Define $h_k = \frac{\partial f_k}{\partial x_k}$. Because $\frac{\partial h_k}{\partial x_k} > 0$ and $\frac{\partial h_k}{\partial y_k} < 0$, it follows that h_k is strictly increasing in x_k and

strictly decreasing in y_k. Consequently, for a fixed y_k, $h_k(x_k, y_k)$ is an invertible function of x_k. Denote by h_k^{-1} the inverse of $h_k(x_k, 0)$. Suppose further that $h_k(x_k, y_k) = h_k(x_k - y_k, 0)$ and finally that $\lim_{x_k \downarrow y_k - \kappa_k} h_k(x_k, y_k) = -\infty$. Now consider the following optimization problem in the variables $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{n} f_k(x_k, y_k) \\ \text{subject to} \quad & x_k,\, y_k \geqslant 0, \\ & \kappa_k + x_k - y_k \geqslant 0 \text{ for } k = 1, \ldots, n, \\ & \text{and } \sum_{k=1}^{n} x_k = \eta,\, \sum_{k=1}^{n} y_k = \theta \text{ for some } \eta, \theta \geqslant 0. \end{aligned}$$

(i) The solution to the problem (x_k^*, y_k^*) depends on two real numbers ψ, ω that satisfy the equality constraints $\sum_k x_k^* = \eta$ and $\sum_k y_k^* = \theta$. The solution exists provided that $\psi \leqslant \omega$. If $\psi < \omega$, then the solution is unique and yields

$$(x_k^*, y_k^*) = \left(\max\left\{0, h_k^{-1}(\psi)\right\}, \max\left\{0, -h_k^{-1}(\omega)\right\}\right).$$

If $\psi = \omega$, then there exists an infinite number of solutions of the form $(x_k^* + \alpha_k, y_k^* + \alpha_k)$ for all $\alpha_k \in \mathbb{R}_+$ meeting the two aforementioned equality constraints.

Without loss of generality, suppose that $h_1(0,0) \leqslant \cdots \leqslant h_n(0,0)$.

(ii) For $\psi < \omega$, consider the following cases:

(a) $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ for some $i = 1, \ldots, j-1$ and $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ for some $j = 2, \ldots, n$.

(b) $h_{j-1}(0,0) \leqslant \omega$ for $j = n+1$ and, either $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ for some $i = 1, \ldots, n-1$ or $h_i(0,0) < \psi$ for $i = n$.

(c) $\psi \leqslant h_{i+1}(0,0)$ for $i = 0$ and, either $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ for some $j = 2, \ldots, n$ or $\omega < h_j(0,0)$ for $j = 1$.

(d) $h_{j-1}(0,0) \leqslant \omega$ for $j = n+1$ and $\psi \leqslant h_{i+1}(0,0)$ for $i = 0$.

In each case, and for the corresponding indexes i and j,

$$\begin{aligned} x_k^* &= \begin{cases} h_k^{-1}(\psi) & , \quad k = 1, \ldots, i \\ 0 & , \quad k = i+1, \ldots, n \end{cases}, \\ y_k^* &= \begin{cases} 0 & , \quad k = 1, \ldots, j-1 \\ -h_k^{-1}(\omega) & , \quad k = j, \ldots, n \end{cases}. \end{aligned}$$

(iii) For $\psi = \omega$, consider the following cases:

(a) either $h_i(0,0) < \psi < h_j(0,0)$ for some $j = 2, \ldots, n$ and $i = j-1$, or $h_i(0,0) < \psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) < h_j(0,0)$ for some $i = 1, \ldots, j-2$ and some $j = 3, \ldots, n$.

(b) for $j = n+1$, either $h_i(0,0) < h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) = \omega$ for some $i = 1, \ldots, j-2$ or $h_{j-1}(0,0) < \omega$ with $i = n$.

(c) for $i = 0$, either $\psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) < h_j(0,0)$ for some $j = 2, \ldots, n$ or $\psi < h_{i+1}(0,0)$ with $j = 1$.

In each case, and for the corresponding indexes i and j,

$$\begin{aligned} x_k^* &= \begin{cases} h_k^{-1}(\psi) + \alpha_k & , \quad k = 1, \ldots, i \\ \alpha_k & , \quad k = i+1, \ldots, n \end{cases}, \\ y_k^* &= \begin{cases} \alpha_k & , \quad k = 1, \ldots, j-1 \\ -h_k^{-1}(\omega) + \alpha_k & , \quad k = j, \ldots, n \end{cases}. \end{aligned}$$

Proof: The proof of statement (i) consists of two steps. In the first step, we show that the optimization problem stated in the lemma is convex; then we apply Karush-Kuhn-Tucker (KKT) conditions to said problem, and finally reformulate these conditions into a reduced number of equations. The bulk of this proof comes later, in the second step, where we proceed to solve the system of equations for the two cases considered in the lemma, $\psi < \omega$ and $\psi = \omega$. Lastly, statements (ii) and (iii) follow from (i).

To see that the problem is convex, simply observe that the objective function is convex on account of $H(f_k) \succeq 0$, and that the inequality and equality constraint functions are affine. Since the objective and constraint functions are also differentiable and Slater's constraint qualification holds, KKT conditions are necessary and sufficient conditions for optimality [41]. Systematic application of these optimality conditions leads to the Lagrangian cost,

$$\mathcal{L} \;=\; \sum f_k(x_k, y_k) \;-\; \sum \lambda_k x_k \;-\; \sum \mu_k y_k \;+\; \sum \nu_k (y_k \;-\; \kappa_k \;-\; x_k) \;-\; \psi \left(\sum x_k - \eta\right) \;+\; \omega \left(\sum y_k - \theta\right),$$

and finally to the conditions

$$\begin{aligned}
&x_k \geqslant 0,\, y_k \geqslant 0,\, \kappa_k + x_k - y_k \geqslant 0,\\
&\textstyle\sum x_k = \eta,\, \sum y_k = \theta, && \text{(primal feasibility)}\\
&\lambda_k \geqslant 0,\, \mu_k \geqslant 0,\, \nu_k \geqslant 0, && \text{(dual feasibility)}\\
&\lambda_k\, x_k = 0,\, \mu_k\, y_k = 0,\\
&\nu_k\,(y_k - \kappa_k - x_k) = 0, && \text{(complementary slackness)}\\
&\tfrac{\partial \mathcal{L}}{\partial x_k} = h_k(x_k, y_k) - \lambda_k - \nu_k - \psi = 0,\\
&\tfrac{\partial \mathcal{L}}{\partial y_k} = h_k(x_k, y_k) + \mu_k - \nu_k - \omega = 0. && \text{(dual optimality)}
\end{aligned}$$

Because $\lim_{x_k \downarrow y_k - \kappa_k} h_k(x_k, y_k) = -\infty$, it follows from the dual optimality conditions that $\kappa_k + x_k - y_k > 0$, which implies, by complementary slackness, that $\nu_k = 0$. Subsequently, we may rewrite the dual optimality conditions as $\lambda_k = h_k(x_k, y_k) - \psi$ and $\mu_k = \omega - h_k(x_k, y_k)$. By eliminating the slack variables λ_k, μ_k, we obtain the simplified conditions $h_k(x_k, y_k) \geqslant \psi$ and $h_k(x_k, y_k) \leqslant \omega$. Lastly, we substitute the above expressions of λ_k and μ_k into the complementary slackness conditions, so that we can formulate the dual optimality and complementary slackness conditions equivalently as

$$h_k(x_k, y_k) \geqslant \psi, \tag{4}$$
$$h_k(x_k, y_k) \leqslant \omega, \tag{5}$$
$$(h_k(x_k, y_k) - \psi)\, x_k = 0, \tag{6}$$
$$(h_k(x_k, y_k) - \omega)\, y_k = 0. \tag{7}$$

In the following, we shall proceed to solve these equations which, together with the primal and dual feasibility conditions, are necessary and sufficient conditions for optimality. To this end, first note that, if $\psi > \omega$, then there exists no (x_k, y_k) that satisfies equations (4) and (5) at the same time, and consequently, as stated in part (i) of the lemma, there is no solution. Concordantly, next we shall study the case when $\psi < \omega$; afterwards we shall tackle the other case when $\psi = \omega$.

Before plunging into the analysis of the former case, recall that the function h_k is strictly increasing in x_k and strictly decreasing in y_k. Having said this, observe that, under the assumption $\psi < \omega$, the variables x_k and y_k cannot be positive simultaneously by virtue of equations (6) and (7). Bearing this in mind, consider these three possibilities for each k: $h_k(0,0) < \psi$, $\psi \leqslant h_k(0,0) \leqslant \omega$ and $\omega < h_k(0,0)$.

When $h_k(0,0) < \psi$, the only conclusion consistent with (4) and with the fact that h_k is strictly increasing in x_k is that $x_k > 0$. Since x_k must be positive, the complementary slackness condition (6) implies that $h_k(x_k, y_k) = \psi$ and, because of (7), that $y_k = 0$. As a result, x_k must satisfy $h_k(x_k, 0) = \psi$, or equivalently, $x_k = h_k^{-1}(\psi)$. Next, we show that the solution $(x_k, 0)$ is unique. For this purpose, suppose that $y_k > 0$ and, in consequence, that $x_k = 0$. It follows from (7), however, that $h_k(0, y_k) = \omega$, which contradicts the fact that h_k is a strictly decreasing function of y_k. In the end, we verify that $x_k = y_k = 0$ does not satisfy (4) and thus prove that $(x_k, y_k) = (h_k^{-1}(\psi), 0)$ is the unique minimizer of the objective function when $h_k(0,0) < \psi$.

Now consider the case when $\psi \leqslant h_k(0,0) \leqslant \omega$. First, suppose that $x_k > 0$, and therefore that $y_k = 0$. By complementary slackness, it follows that $h_k(x_k, 0) = \psi$, which is not consistent with the fact that h_k is strictly increasing in x_k. Consequently, x_k cannot be positive. Secondly, assume that x_k is zero and y_k positive. Under this assumption, equation (7) implies that $h_k(0, y_k) = \omega$, a contradiction since h_k is a strictly decreasing function of y_k. Accordingly, y_k cannot be positive either. Finally, check that $x_k = y_k = 0$ satisfies the optimality conditions and hence it is the unique solution.

The last possibility corresponds to the case when $\omega < h_k(0,0)$. Note that, in this case, the only conclusion consistent with (5) and with the fact that h_k is strictly decreasing in y_k is that $y_k > 0$. Thus, because of (7), y_k must satisfy $h_k(0, y_k) = \omega$. Recalling from the lemma that $h_k(x_k, y_k) = h_k(x_k - y_k, 0)$, we may express the condition $h_k(0, y_k) = \omega$ equivalently as $y_k = -h_k^{-1}(\omega)$. Lastly, we check that this solution is unique in the case under study. To this end, note that a solution such that $x_k > 0$ and $y_k = 0$ contradicts the fact that h_k is strictly increasing in x_k. As a result, x_k cannot be positive. Finally, we confirm that equation (5) does not hold for $x_k = y_k = 0$ and therefore prove that $(x_k, y_k) = (0, -h_k^{-1}(\omega))$ is the unique solution when $\omega < h_k(0,0)$.

In summary, $x_k = h_k^{-1}(\psi)$ if $h_k(0,0) < \psi$, or equivalently, $h_k^{-1}(\psi) > 0$; otherwise $x_k = 0$. Further, $y_k = -h_k^{-1}(\omega)$ if $h_k(0,0) > \omega$, or equivalently, $h_k^{-1}(\omega) < 0$; otherwise $y_k = 0$. Accordingly, we may write the solution compactly as

$$(x_k, y_k) = \left(\max\left\{0, h_k^{-1}(\psi)\right\}, \max\left\{0, -h_k^{-1}(\omega)\right\}\right),$$

where ψ, ω must satisfy the primal equality constraints $\sum_k x_k = \eta$ and $\sum_k y_k = \theta$.

Having examined the case when $\psi < \omega$, next we proceed to solve the optimality conditions at hand for $\psi = \omega$. Observe that, in this new case, (4) and (5) transform into the equation

$$h_k(x_k, y_k) = \psi. \tag{8}$$

Moreover, note that any pair (x_k, y_k) satisfying (8) also meets the complementary slackness conditions (6) and (7). However, notice that this does not mean that all those pairs are optimal. To elaborate on this point, consider the following three possibilities for each k: $h_k(0,0) < \psi$, $h_k(0,0) = \psi$ and $\psi < h_k(0,0)$.

In the case when $h_k(0,0) < \psi$, the only condition consistent with (8) and with the fact that h_k is strictly increasing in x_k is that $x_k > 0$. From the lemma, it is immediate that $\frac{\partial h_k}{\partial x_k} = -\frac{\partial h_k}{\partial y_k}$, which implies that x_k must also be greater than y_k. Hence, the set of solutions is

$$\{(x_k, y_k)\colon h_k(x_k, y_k) = \psi,\, x_k > y_k\},$$

where every pair in this set must also fulfill the primal equality conditions. Let x'_k satisfy $h_k(x'_k, 0) = \psi$, or equivalently, $x'_k = h_k^{-1}(\psi)$. Then, because $h_k(x'_k + \alpha_k, \alpha_k) = \psi$ for any $\alpha \geqslant 0$, this set may be recast equivalently as

$$\{(x_k, y_k)\colon x_k = x'_k + \alpha_k,\, y_k = \alpha_k\}.$$

For the two remaining cases, i.e., $h_k(0,0) = \psi$ and $\psi < h_k(0,0)$, the set of solutions is obtained in a completely analogous way as above. In the former case, the pairs (x_k, y_k) must satisfy $x_k = y_k$, and the set of solutions may be expressed as

$$\{(x_k, y_k)\colon x_k = \alpha_k,\, y_k = \alpha_k\}.$$

In the latter case, it follows that $y_k > x_k$ and, consequently, that the set of solutions is

$$\{(x_k, y_k)\colon x_k = \alpha_k,\, y_k = y'_k + \alpha_k\},$$

where y'_k must satisfy $h_k(0, y'_k) = \psi$.

To sum up, the case $\psi = \omega$ leads to the following solutions: $x_k = h_k^{-1}(\psi) + \alpha_k$ if $h_k(0,0) < \psi$, or equivalently, $h_k^{-1}(\psi) > 0$; otherwise $x_k = \alpha_k$. In addition, $y_k = -h_k^{-1}(\omega) + \alpha_k$ if $h_k(0,0) > \omega$, or equivalently, $h_k^{-1}(\omega) < 0$; otherwise $y_k = \alpha_k$. Accordingly, the solutions (x_k, y_k) yield

$$\left(\max\left\{0, h_k^{-1}(\psi)\right\} + \alpha_k, \max\left\{0, -h_k^{-1}(\omega)\right\} + \alpha_k\right), \tag{9}$$

for some ψ, ω and nonnegative sequence $\alpha_1, \ldots, \alpha_n$ such that $\sum_k x_k = \eta$ and $\sum_k y_k = \theta$. Note that, although $\psi = \omega$, we intentionally write ω instead of ψ to highlight that the solutions for $\psi < \omega$ and for $\psi = \omega$ just differ in the term α_k, as we claimed in part (i) of the lemma.

To complete the proof of statement (i), it suffices to show that the number of solutions is infinite when $\psi = \omega$. To this end, simply observe that there exists an infinite number of sequences $\alpha_1, \ldots, \alpha_n$ such that

$$\sum_k x_k = \sum_k h_k^{-1}(\psi) + \sum_k \alpha_k = \eta \quad \text{and}$$

$$\sum_k y_k = -\sum_k h_k^{-1}(\psi) + \sum_k \alpha_k = \theta,$$

which results in an infinite number of solutions of the form given in (9).

Now we proceed to prove (ii), which is an immediate consequence of (i). For this purpose, observe that if $\psi \leqslant h_{i+1}(0,0) \leqslant \cdots \leqslant h_n(0,0)$ holds for some $i = 0, \ldots, n-1$, then $h_{i+1}^{-1}(\psi), \ldots, h_n^{-1}(\psi) \leqslant 0$, and accordingly $x_{i+1} = \cdots = x_n = 0$. Similarly, if $h_1(0,0) \leqslant \cdots \leqslant h_{j-1}(0,0) \leqslant \omega$ is satisfied for some $j = 2, \ldots, n+1$, then $h_1^{-1}(\omega), \ldots, h_{j-1}^{-1}(\omega) \geqslant 0$, and thus $y_1 = \cdots = y_{j-1} = 0$.

Note that the particular case when the index i ranges from 1 to $j-1$ and the index j goes from 2 to n is the case described in (ii) (a), which corresponds to $\eta, \theta > 0$. Further, observe that the case assumed in (ii) (b), i.e., when $j = n+1$, implies that $\theta = 0$. Here, the index i starts at 1, therefore excluding $\eta = 0$, and ends at n, including the possibility that $x_i > 0$ for all i. In part (ii) (c), we consider $i = 0$, which is equivalent to the condition $\eta = 0$. In this case, the index j starts at 1, permitting $y_j > 0$ for all j, and ends at n, avoiding $\theta = 0$. Finally, the case described in (ii) (d), namely when $j = n+1$ and $i = 0$, is precisely the trivial case $x = y = 0$.

In order to verify statement (iii), we proceed analogously by noting that if $\psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0)$ holds for some $i = 1, \ldots, j-2$ and some $j = 3, \ldots, n$, then $h_{i+1}^{-1}(\psi) = \cdots = h_{j-1}^{-1}(\psi) = 0$, and consequently $x_k = y_k = \alpha_k$ for $k = i+1, \ldots, j-1$. ■

The previous lemma presented the solution to a resource allocation problem that minimizes a rather general but convex objective function, subject to affine constraints. Our next theorem, Theorem 3, applies the results of this lemma to the special case of the objective function of problem (1). In doing so, we shall confirm the intuition that there must exist a set of ordered pairs (ρ, σ) where the privacy risk vanishes and another set where it does not. We shall refer to the former set as the *critical-privacy region* and formally define it as

$$\mathscr{C} = \{(\rho, \sigma)\colon \mathcal{R}(\rho, \sigma) = 0\}.$$

The latter set will be the complementary set $\bar{\mathscr{C}}$ and we shall refer to it as the *noncritical-privacy region*.

Before proceeding with Theorem 3, first we shall introduce what we term *forgery* and *suppression thresholds*, two sequences of rates that will play a fundamental role in the characterization of the solution to the minimization problem defining the privacy-forgery-suppression function. Secondly, we shall investigate certain properties of these thresholds in Proposition 2. And thereafter, we shall introduce some definitions that will facilitate the exposition of the aforementioned theorem.

Let $Q_i = \sum_{k=1}^{i} q_k$ and $P_i = \sum_{k=1}^{i} p_k$ be the cumulative distribution functions corresponding to q and p. Denote by $\bar{Q}_i = \sum_{k=i}^{n} q_k$ and $\bar{P}_i = \sum_{k=i}^{n} p_k$ the complementary cumulative distribution functions of q and p. Define the *forgery thresholds* ρ_i as

$$\rho_i = \begin{cases} P_i \frac{q_i}{p_i} - Q_i & , \quad i = 1, \ldots, j-1 \\ \frac{P_{j-1}}{\bar{P}_j}(\bar{Q}_j - \sigma) - Q_{j-1}, & i = j \\ \infty & , \quad i = j+1 \end{cases},$$

for $j = 2, \ldots, n$. Additionally, define the *suppression thresholds* σ_j as

$$\sigma_j = \bar{Q}_j - \bar{P}_j \frac{q_j}{p_j}$$

for $j = 1, \ldots, n$, and $\sigma_0 = 1$. Observe that $\rho_1 = \sigma_n = 0$ and that the forgery threshold ρ_j is a linear function of σ. We shall refer to this latter threshold as the *critical forgery-suppression threshold* and denote it also by $\rho_{\text{crit}}(\sigma)$. The reason is that said threshold will determine the boundary of the critical-privacy region, as we shall see later. The following result, Proposition 2, characterizes the monotonicity of the forgery and the suppression thresholds.

Proposition 2 (Monotonicity of Thresholds):

(i) For $j = 3, \ldots, n$ and $i = 1, \ldots, j-2$, the forgery thresholds satisfy $\rho_i \leqslant \rho_{i+1}$, with equality if, and only if, $\frac{q_i}{p_i} = \frac{q_{i+1}}{p_{i+1}}$.

(ii) For $j = 2, \ldots, n$, the suppression thresholds satisfy $\sigma_j \leqslant \sigma_{j-1}$, with equality if, and only if, $\frac{q_j}{p_j} = \frac{q_{j-1}}{p_{j-1}}$.

(iii) Further, for any $j = 2, \ldots, n$ and any $\sigma \in (\sigma_j, \sigma_{j-1}]$, the critical forgery-suppression threshold satisfies $\rho_j(\sigma) \geqslant \rho_{j-1}$, with equality if, and only if, $\sigma = \sigma_{j-1}$.

Proof: The first statement can be shown from the definition of the forgery thresholds by routine algebraic manipulation and under the labeling assumption (3). To this end, it is helpful to note that

$$P_i \frac{q_{i+1}}{p_{i+1}} - Q_i = P_{i+1} \frac{q_{i+1}}{p_{i+1}} - Q_{i+1}.$$

The second statement can be shown analogously, observing that

$$\bar{Q}_j - \bar{P}_j \frac{q_{j-1}}{p_{j-1}} = \bar{Q}_{j-1} - \bar{P}_{j-1} \frac{q_{j-1}}{p_{j-1}}.$$

For the last statement, use the definitions of the forgery and the suppression thresholds to note that the condition $\rho_j(\sigma) \geqslant \rho_{j-1}$ is equivalent to $\sigma \leqslant \sigma_{j-1}$. ∎

Prior to investigate a closed-form solution to the problem (1), we introduce some definitions for ease of presentation. For $i = 1, \ldots, j-1$ and $j = 2, \ldots, n$, define

$$\begin{aligned} \tilde{q} &= \big(\, Q_i \, , \, q_{i+1} \, , \ldots , \, q_{j-1} \, , \, \bar{Q}_j \, \big), \\ \tilde{r} &= \big(\, \rho \, , \quad 0 \quad , \ldots , \quad 0 \quad , \, 0 \, \big), \\ \tilde{s} &= \big(\, 0 \, , \quad 0 \quad , \ldots , \quad 0 \quad , \, \sigma \, \big), \\ \tilde{p} &= \big(\, P_i \, , p_{i+1} \, , \ldots , p_{j-1} \, , \, \bar{P}_j \, \big), \end{aligned}$$

where $\tilde{q}$ and $\tilde{p}$ are distributions in the probability simplex of $j-i+1$ dimensions, and $\tilde{r}$ and $\tilde{s}$ are tuples of the same dimension that represent a forgery strategy and a suppression strategy, respectively. Particularly, note that the indexes $i = 1$ and $j = n$ lead to $\tilde{q} = q$ and $\tilde{p} = p$.

Theorem 3: Let $\partial\mathscr{C}$ be the boundary of $\mathscr{C}$, and $\operatorname{cl}\bar{\mathscr{C}}$ the closure of $\bar{\mathscr{C}}$.

(i) $\partial\mathscr{C} \subset \mathscr{C}$ and

$$\partial\mathscr{C} = \{(\rho, \sigma) \colon \rho = \rho_j(\sigma), \sigma \in [\sigma_j, \sigma_{j-1}], \text{ for } j = 2, \ldots, n\}.$$

(ii) For any $(\rho, \sigma) \in \operatorname{cl}\bar{\mathscr{C}}$, either $\rho \in [\rho_i, \rho_{i+1}]$ for $i = 1$ or $\rho \in (\rho_i, \rho_{i+1}]$ for some $i = 2, \ldots, j-1$, and either $\sigma \in [\sigma_j, \sigma_{j-1}]$ for $j = n$ or $\sigma \in (\sigma_j, \sigma_{j-1}]$ for some $j = 2, \ldots, n-1$. Then, for the corresponding indexes i, j, the optimal forgery and suppression strategies are

$$r_k^* = \begin{cases} \frac{p_k}{P_i}(Q_i + \rho) - q_k \, , & k = 1, \ldots, i \\ 0 \qquad\qquad\quad , & k = i+1, \ldots, n \end{cases},$$

$$s_k^* = \begin{cases} 0 \qquad\qquad\quad , & k = 1, \ldots, j-1 \\ q_k - \frac{p_k}{\bar{P}_j}(\bar{Q}_j - \sigma), & k = j, \ldots, n \end{cases},$$

and the corresponding, minimum KL divergence yields the privacy-forgery-suppression function

$$\mathcal{R}(\rho,\sigma) = \mathrm{D}\left(\frac{\tilde{q}+\tilde{r}-\tilde{s}}{1+\rho-\sigma}\middle\|\,\tilde{p}\right).$$

Proof: The proof is structured as follows. We begin by showing that the optimization problem (1) may be construed as a particular case of that stated in Lemma 1. Accordingly, we apply this lemma, namely the cases (ii) and (iii), to obtain the optimal forgery and suppression strategies. The application of the former case allows us to derive the solution for $(\rho,\sigma)\in\bar{\mathscr{C}}$. The latter case enables us, first, to confirm that this solution is also valid on $\partial\bar{\mathscr{C}}$, and secondly, to prove statement (i). Lastly, we complete the proof of (ii) by expressing function (1) in terms of the optimal apparent distribution.

Use the definition of KL divergence to write the objective function of the optimization problem as $\mathrm{D}(t\,\|\,p)=\sum_k t_k\log\frac{t_k}{p_k}$, with $t=\frac{q+r-s}{1+\rho-\sigma}$. Observe that the functions $f_k(r_k,s_k)=t_k\log\frac{t_k}{p_k}$ are twice differentiable on $\{(r_k,s_k)\colon q_k+r_k-s_k>0\}$. Denote by h_k the derivative of f_k with respect to r_k,

$$h_k(r_k,s_k)=\frac{1}{1+\rho-\sigma}\left(\log\frac{q_k+r_k-s_k}{(1+\rho-\sigma)p_k}+1\right). \tag{10}$$

Then, note that the functions f_k and h_k satisfy the assumptions of Lemma 1, and that the inequality and equality constraints of function (1) coincide with those in the lemma. This exposes the structure of the optimization problem as a special case of the resource allocation lemma.

Before proceeding any further, notice from (10) that $h_k(r_k,0)$ is a strictly increasing function of r_k and hence invertible. Note also that, according to the lemma, the solutions are completely determined by the inverse of this function, which is denoted by h_k^{-1} and yields

$$h_k^{-1}(\phi)=p_k(1+\rho-\sigma)2^{(1+\rho-\sigma)\phi-1}-q_k.$$

Finally, observe that the assumption $h_1(0,0)\leqslant\cdots\leqslant h_n(0,0)$ in the lemma is equivalent to the labeling assumption (3), as $h_k(0,0)$ is a strictly increasing function of $\frac{q_k}{p_k}$.

Next we apply Lemma 1 (ii), where it is assumed the condition $\psi<\omega$. We start with case (ii) (a). On account of part (i) of the lemma, the optimal forgery strategy must satisfy

$$\rho=\sum_{k=1}^{i}h_k^{-1}(\psi)=P_i(1+\rho-\sigma)2^{(1+\rho-\sigma)\psi-1}-Q_i,$$

or equivalently,

$$\psi=\frac{1}{1+\rho-\sigma}\left(\log\frac{Q_i+\rho}{(1+\rho-\sigma)P_i}+1\right).$$

Analogously for the suppression strategy,

$$\sigma=-\sum_{k=j}^{n}h_k^{-1}(\omega)=\bar{Q}_j-\bar{P}_j(1+\rho-\sigma)2^{(1+\rho-\sigma)\omega-1},$$

and therefore

$$\omega=\frac{1}{1+\rho-\sigma}\left(\log\frac{\bar{Q}_j-\sigma}{(1+\rho-\sigma)\bar{P}_j}+1\right).$$

Then it suffices to substitute the expressions of ψ and ω into the function h_k^{-1}, to obtain the nonzero optimal solutions claimed in assertion (ii) of the theorem.

Now we proceed to confirm the interval of values of ρ and σ where these solutions are defined. In the case under study, ψ and ω satisfy $h_i(0,0)<\psi\leqslant h_{i+1}(0,0)$ for some $i=1,\ldots,j-1$ and $h_{j-1}(0,0)\leqslant\omega<h_j(0,0)$ for some $j=2,\ldots,n$. We split the discussion into two cases, namely $i<j-1$ and $i=j-1$.

Assume the former case. Observe that the condition $h_i(0,0)<\psi$ is equivalent to

$$\frac{1}{1+\rho-\sigma}\left(\log\frac{q_i}{(1+\rho-\sigma)p_i}+1\right)<\frac{1}{1+\rho-\sigma}\left(\log\frac{Q_i+\rho}{(1+\rho-\sigma)P_i}+1\right)$$

and finally, after routine algebraic manipulation, to

$$\rho>P_i\frac{q_i}{p_i}-Q_i.$$

Similarly, the upper-bound condition $\psi\leqslant h_{i+1}(0,0)$ leads to

$$\rho\leqslant P_i\frac{q_{i+1}}{p_{i+1}}-Q_i.$$

Hence, the intervals resulting from imposing $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ are of the form $(\rho_i, \rho_{i+1}]$. The monotonicity of the thresholds ρ_i, demonstrated in Proposition 2, guarantees that these intervals are contiguous and nonoverlapping. In an analogous manner, it can be shown that the condition $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ leads to intervals of the form $(\sigma_j, \sigma_{j-1}]$, also contiguous and nonoverlapping by virtue of Proposition 2.

Now assume the latter case, where $h_i(0,0) < \psi < \omega < h_j(0,0)$ with $i = j-1$. On the one hand, the assumption $h_{j-1}(0,0) < \psi$ is, as shown above, equivalent to the condition $\rho > \rho_{j-1}$. On the other hand, straightforward manipulation allows us to write the inequality $\psi < \omega$ as

$$\rho < \frac{P_{j-1}}{\bar{P}_j}(\bar{Q}_j - \sigma) - Q_{j-1}.$$

Combining these two bounds on ψ, we obtain the interval $(\rho_{j-1}, \rho_{\text{crit}}(\sigma))$. With this last interval, we complete the range of validity of the solution for the case (ii) (a) in the lemma. Ultimately, it is easy to verify that, in those intervals of ρ and σ, the optimal apparent profile $t = \frac{q+r-s}{1+\rho-\sigma}$ does not coincide with the population's profile p. In consequence, $\mathrm{D}(t \,\|\, p) > 0$.

Next, we turn to case (ii) (b) of the lemma. Here, the assumption $h_n(0,0) \leqslant \omega$ leads to $\sigma = 0$, or equivalently, to the solution $s = 0$. Note that, precisely, this is the solution given in the theorem for $\sigma = \sigma_j$ with $j = n$. On the other hand, the application of the condition $\sum_{k=1}^{i} r_k = \rho$ results in the same optimal forgery strategy obtained in case (ii) (a). Proceeding analogously as in this case, from the assumptions on ψ we derive the intervals of values of ρ where the solution is defined: $(\rho_i, \rho_{i+1}]$ for $i = 1, \ldots, n-1$ and (ρ_i, ρ_{i+1}) for $i = n$. Given these intervals, it is then straightforward to check that $\mathcal{R}(\rho, 0) = 0$ if, and only if, $\rho \geqslant \rho_n$. This provides us with the pairs $(\rho, 0)$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$.

In case (ii) (c), the condition $\psi \leqslant h_1(0,0)$ means that $\rho = 0$, or equivalently, $r = 0$. Observe that this is the solution stated in the theorem for $\rho = \rho_i$ with $i = 1$. Then again, the condition $\sum_{k=j}^{n} s_k = \sigma$ leads to the same optimal suppression strategy found in case (ii) (a). From the assumptions in the lemma on ω, we obtain the intervals $(\sigma_j, \sigma_{j-1}]$ for $j = 2, \ldots, n$ and (σ_j, σ_{j-1}) for $j = 1$. Then, we verify that $\mathcal{R}(0, \sigma) = 0$ if, and only if, $\sigma \geqslant \sigma_1$, from which it follows the pairs $(0, \sigma)$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$.

Finally, the case (ii) (d) in the lemma, in which $h_n(0,0) \leqslant \omega$ and $\psi \leqslant h_1(0,0)$, corresponds to the trivial case $\sigma = \sigma_j$ for $j = n$ and $\rho = \rho_i$ for $i = 1$, that is, the solution $r = s = 0$.

After having applied Lemma 1 (ii) to function (1), now we proceed with case (iii) (a). In applying it, we shall show that the solution claimed in the theorem is also valid for the extreme values of the intervals in case (ii) (a), specifically the set

$$\{(\rho, \sigma) \colon \rho = \rho_{\text{crit}}(\sigma), \sigma \in (\sigma_j, \sigma_{j-1}] \text{ for } j = 3, \ldots, n, \text{ and } \sigma \in (\sigma_j, \sigma_{j-1}) \text{ for } j = 2\}.$$

Assume the case (iii) (a) in which $h_i(0,0) < \psi = \omega < h_j(0,0)$ for some $j = 2, \ldots, n$ and $i = j-1$. Under this assumption, the equality constraint $\sum_{k=1}^{i} r_k = \rho$ in the lemma is equivalent, after simple algebraic manipulation, to

$$\psi = \frac{1}{1+\rho-\sigma}\left(\log \frac{Q_{j-1} + \rho - \zeta}{(1+\rho-\sigma)P_{j-1}} + 1\right), \tag{11}$$

where we define $\zeta = \sum_{k=1}^{n} \alpha_k$. Similarly, the equality constraint $\sum_{k=j}^{n} s_k = \sigma$ becomes

$$\omega = \frac{1}{1+\rho-\sigma}\left(\log \frac{\bar{Q}_j - \sigma + \zeta}{(1+\rho-\sigma)\bar{P}_j} + 1\right).$$

But $\psi = \omega$, therefore

$$\frac{Q_{j-1} + \rho - \zeta}{P_{j-1}} = \frac{\bar{Q}_j - \sigma + \zeta}{\bar{P}_j},$$

or equivalently,

$$\rho = \rho_{\text{crit}}(\sigma) + \frac{\zeta}{\bar{P}_j}.$$

In short, the assumption $\psi = \omega$ imposes the condition $(\rho, \sigma) \succeq (\rho_{\text{crit}}(\sigma), \sigma)$ for some nonnegative sequence $\alpha_1, \ldots, \alpha_n$ satisfying the above equality. Next we examine, for a given σ, these two possibilities, $\rho = \rho_{\text{crit}}(\sigma)$ and $\rho > \rho_{\text{crit}}(\sigma)$.

Consider the former possibility and observe that $\rho = \rho_{\text{crit}}(\sigma)$ if, and only if, $\alpha_k = 0$ for $k = 1, \ldots, n$. According to the lemma, the nonzero optimal solutions yield

$$\begin{aligned} r_k = h_k^{-1}(\psi) &= p_k \frac{Q_{j-1} + \rho_{\text{crit}}(\sigma)}{P_{j-1}} - q_k \\ &= p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma) - q_k \end{aligned}$$

for $k = 1, \ldots, j-1$, and

$$s_k = -h_k^{-1}(\psi) = q_k - p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma)$$

for $k = j, \ldots, n$, that is, the solutions obtained after applying case (ii) (a), but evaluated at $\rho = \rho_{\text{crit}}(\sigma)$. From these expression for r and s, it is immediate to verify then that $t = p$ and thus $\mathcal{R}(\rho, \sigma) = 0$.

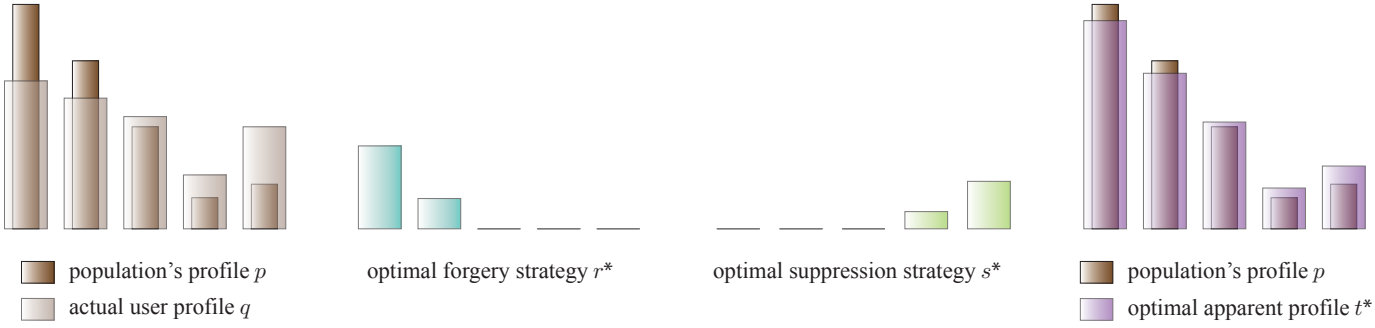


Fig. 2: A user's item distribution is perturbed according to two optimal forgery and suppression strategies, in order for the resulting profile to minimize the KL divergence with respect to the population's distribution.

Now we assume the latter possibility, i.e., $(\rho, \sigma) \succ (\rho_{\text{crit}}(\sigma), \sigma)$, to show that the privacy-risk function also vanishes for these values of ρ and σ. On account of part (iii) (a) of the lemma and (11), we derive the optimal forgery and suppression strategies

$$r_k = p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma) + \frac{p_k\,\zeta}{\bar{P}_j} - q_k + \alpha_k$$

and $s_k = \alpha_k$ for $k = 1, \ldots, j-1$, and

$$s_k = q_k - p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma) - \frac{p_k\,\zeta}{\bar{P}_j} + \alpha_k$$

and $r_k = \alpha_k$ for $k = j, \ldots, n$. Then, we substitute r and s back into the apparent profile t and check that $\mathrm{D}(t \,\|\, p) = 0$. In doing so, we determine the pairs $(\rho, \sigma) \succ 0$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$, and finally obtain the expression for the boundary of the critical-privacy region claimed in statement (i) of the theorem.

To conclude the proof, it remains only to write the privacy-risk function $\mathcal{R}(\rho, \sigma) = \sum_{k=1}^{n} t_k \log \frac{t_k}{p_k}$ in terms of the optimal apparent distribution. With this aim, we split the summation into three parts. The first part, corresponding to $t_k = \frac{p_k(Q_i+\rho)}{P_i(1+\rho-\sigma)}$, is

$$\sum_{k=1}^{i} t_k \log \frac{t_k}{p_k} = \frac{Q_i + \rho}{1 + \rho - \sigma} \log \frac{Q_i + \rho}{(1 + \rho - \sigma) P_i},$$

where we leverage on the fact that $\frac{t_k}{p_k}$ does not depend on k. The second part of the sum, corresponding to $t_k = \frac{q_k}{1+\rho-\sigma}$, yields

$$\sum_{k=i+1}^{j-1} t_k \log \frac{t_k}{p_k} = \sum_{k=i+1}^{j-1} \frac{q_k}{1 + \rho - \sigma} \log \frac{q_k}{(1 + \rho - \sigma) p_k}.$$

The last part, corresponding to $t_k = \frac{p_k(\bar{Q}_j-\sigma)}{\bar{P}_j(1+\rho-\sigma)}$, is

$$\sum_{k=j}^{n} t_k \log \frac{t_k}{p_k} = \frac{\bar{Q}_j - \sigma}{1 + \rho - \sigma} \log \frac{\bar{Q}_j - \sigma}{(1 + \rho - \sigma) \bar{P}_j},$$

where we also note that $\frac{t_k}{p_k}$ does not depend on k either. Now, it is straightforward to identify the terms of $\mathcal{R}(\rho, \sigma)$ as the KL divergence between the distributions

$$\left(\frac{Q_i + \rho}{1 + \rho - \sigma}, \frac{q_{i+1}}{1 + \rho - \sigma}, \ldots, \frac{q_{j-1}}{1 + \rho - \sigma}, \frac{\bar{Q}_j - \sigma}{1 + \rho - \sigma} \right)$$

and

$$\left(P_i, p_{i+1}, \ldots, p_{j-1}, \bar{P}_j \right),$$

precisely the distributions stated in the theorem. ■

In light of Theorem 3, we would like to remark the intuitive principle that both the optimal forgery and suppression strategies follow. On the one hand, the forgery strategy suggests adding ratings to those categories with a low ratio $\frac{q_k}{p_k}$, that is, to those in which the user's interest is considerably lower than the population's. On the other hand, the suppression strategy recommends eliminating ratings from those categories where the ratio $\frac{q_k}{p_k}$ is high, i.e., where the interest of the user exceeds that of the population.

Another straightforward consequence of Theorem 3 is the role of the forgery and the suppression thresholds. In particular, we identify ρ_i as the forgery rate beyond which the components of r_k for $k = 1, \ldots, i$ become positive. A similar reasoning applies to σ_j, which indicates the suppression rate beyond which the components of s_k for $k = j, \ldots, n$ are positive. In a nutshell, these thresholds determine the number of nonzero components of the optimal strategies.

Also, from this theorem we deduce that the perturbation of the user profile does not *only* affect those categories where either $r_k > 0$ or $s_k > 0$. In fact, since we are dealing with relative frequencies, the components of the apparent distribution t_k belonging to the categories $k = i+1, \ldots, j-1$ are normalized by $\frac{1}{1+\rho-\sigma}$. Fig. 2 illustrates these three conclusions by means of a simple example with $n = 5$ categories of interest.

In this example we consider a user who is disposed to submit a percentage of false ratings $\rho \in (\rho_2, \rho_3]$, and to refrain from sending a fraction of genuine ratings $\sigma \in (\sigma_4, \sigma_3]$. Given these rates, the optimal forgery strategy recommends that the user forge ratings belonging to the categories 1 and 2, where clearly there is a lack of interest, compared to the reference distribution. On the contrary, the suppression strategy specifies that the user eliminate ratings from the categories 4 and 5, that is, from those categories where they show too much interest, again compared to the population's profile. In adopting these two strategies, the apparent user profile approaches the population's distribution, especially in those components where the ratio $\frac{q_k}{p_k}$ deviates significantly from 1. Finally, the component of the apparent profile t_3, which is not directly affected by the forgery and the suppression strategies, gets closer to p_3 as a result of the aforementioned normalization.

In the following subsections, we shall analyze a number of important consequences of Theorem 3.

B. Orthogonality, Continuity and Proportionality

In this subsection we study some interesting properties of the closed-form solution obtained in Sec. IV-A. Specifically, we investigate the orthogonality and continuity of the optimal forgery and suppression strategies, and then establish a proportionality relationship between the optimal apparent user profile and the population's distribution.

Corollary 4 (Orthogonality and Continuity):

(i) For any $(\rho, \sigma) \in \mathrm{cl}\,\bar{\mathscr{C}}$, the optimal forgery and suppression strategies satisfy $r_k^* s_k^* = 0$ for $k = 1, \ldots, n$.

(ii) The components of r^* and s^*, interpreted as functions of ρ and σ respectively, are continuous on $\mathrm{cl}\,\bar{\mathscr{C}}$.

Proof: The proof of (i) is trivial from Theorem 3. To prove statement (ii) we also resort to this theorem. According to it, each component r_k^* may be regarded as a piecewise function of ρ defined on the contiguous, nonoverlapping intervals $[\rho_i, \rho_{i+1}]$ for $i = 1$ and $(\rho_i, \rho_{i+1}]$ for $i = 2, \ldots, j-1$. A direct verification shows that, for any $k = j, \ldots, n$, the component r_k^* is identically zero on the whole interval $[\rho_1, \rho_j]$ and hence continuous. For any $k = 1, \ldots, j-1$, we immediately check the continuity of r_k^* on the interior of each of the intervals parameterized by i. Now we examine the endpoints of such intervals. The continuity at the extreme points ρ_1 and ρ_j is verified straightforwardly as the intervals are closed at these points. Then, we check that the limit at the remaining endpoints ρ_i exists, since

$$\begin{aligned}\lim_{\rho \to \rho_i^-} r_k^*(\rho) &= \frac{p_k}{P_{i-1}}(Q_{i-1} + \rho_i) - q_k \\ &= \frac{p_k}{P_i}(Q_i + \rho_i) - q_k = \lim_{\rho \to \rho_i^+} r_k^*(\rho),\end{aligned}$$

for $i = 2, \ldots, j-1$. Because each limit coincides with the corresponding value $r_k^*(\rho_i)$, we prove the continuity of the components $r_1, \ldots, r_{j-1}$. The proof of the continuity of the components of s^* is analogous to that of r^*. ■

The orthogonality of the optimal forgery and suppression strategies, in the sense indicated by Corollary 4 (i), conforms to intuition—it would not make any sense to submit false ratings to items of a particular category and, at the same time, eliminate genuine ratings from this category. This intuitive result is illustrated in Fig. 2. The second part of Corollary 4 is applied to show our next result, Proposition 5.

Proposition 5 (Proportionality): Define the piecewise functions $\phi(\rho, \sigma) = \frac{Q_i + \rho}{(1+\rho-\sigma)P_i}$ and $\chi(\rho, \sigma) = \frac{\bar{Q}_j - \sigma}{(1+\rho-\sigma)\bar{P}_j}$ on the intervals $[\sigma_j, \sigma_{j-1}]$ for $j = 2, \ldots, n$ and $[\rho_i, \rho_{i+1}]$ for $i = 1, \ldots, j-1$.

(i) For any $j = 2, \ldots, n$ and $i = 1, \ldots, j-1$, and for any $\sigma \in [\sigma_j, \sigma_{j-1}]$ and $\rho \in [\rho_i, \rho_{i+1}]$, the optimal apparent profile t^* and the population's distribution p satisfy

$$\frac{t_1^*}{p_1} = \cdots = \frac{t_i^*}{p_i} = \phi(\rho, \sigma),$$

$$\frac{t_j^*}{p_j} = \cdots = \frac{t_n^*}{p_n} = \chi(\rho, \sigma),$$

and

$$\phi(\rho, \sigma) \leqslant \frac{t_{i+1}^*}{p_{i+1}} \leqslant \cdots \leqslant \frac{t_{j-1}^*}{p_{j-1}} \leqslant \chi(\rho, \sigma).$$

(ii) The function ϕ is continuous and strictly *increasing* in each of its arguments, and satisfies $\phi(\rho, \sigma) \leqslant 1$, with equality if, and only if, $(\rho, \sigma) = (\rho_j(\sigma), \sigma)$.

(iii) The function χ is continuous and strictly *decreasing* in each of its arguments, and satisfies $\chi(\rho, \sigma) \geqslant 1$, with equality if, and only if, $(\rho, \sigma) = (\rho_j(\sigma), \sigma)$.

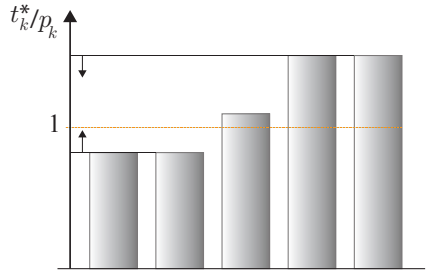


Fig. 3: Proportionality relationship between the optimal user's apparent item distribution and the population's profile. In this figure we show the ratios $\frac{t_k^*}{p_k}$ of the example illustrated in Fig. 2, where the number of categories is $n = 5$, $\rho \in [\rho_2, \rho_3]$ and $\sigma \in [\sigma_4, \sigma_3]$.

Proof: The continuity of the components of t^* on $\mathrm{cl}\,\bar{\mathscr{C}}$ follows from Corollary 4 (ii). This allows us to write the intervals in Theorem 3 as $[\rho_i, \rho_{i+1}]$ and $[\sigma_j, \sigma_{j-1}]$, in lieu of $(\rho_i, \rho_{i+1}]$ and $(\sigma_j, \sigma_{j-1}]$, respectively. From the expressions of r_k^* and s_k^* in the theorem, it is immediate to identify the ratios $\frac{t_k^*}{p_k}$ as either $\phi(\rho, \sigma)$ or $\chi(\rho, \sigma)$. The inner inequalities in statement (i) of this proposition also follow immediately from the labeling assumption (3). Direct manipulation shows that the outer inequalities $\frac{t_i^*}{p_i} \leqslant \frac{t_{i+1}^*}{p_{i+1}}$ and $\frac{t_{j-1}^*}{p_{j-1}} \leqslant \frac{t_j^*}{p_j}$ are equivalent to $\rho \leqslant \rho_{i+1}$ and $\sigma \leqslant \sigma_{j-1}$, respectively. This proves (i).

Next, we proceed to demonstrate the strict monotonicity of ϕ. A simple calculation shows that

$$\frac{\partial \phi}{\partial \rho} = \frac{\bar{Q}_{i+1} - \sigma}{(1 + \rho - \sigma)^2 P_i}.$$

To prove that $\frac{\partial \phi}{\partial \rho} > 0$, it is sufficient to verify that $\bar{Q}_j > \sigma_{j-1}$, or equivalently, that $\bar{P}_j \frac{q_{j-1}}{p_{j-1}} > 0$. Then, by the positivity assumption (2), we immediately see that this latter inequality holds for any $j = 2, \ldots, n$. The strict monotonicity of ϕ in σ also follows from assumption (2).

To complete (ii), we write the condition $\phi(\rho, \sigma) \leqslant 1$ as

$$\rho \leqslant \frac{(1 - \sigma) P_i - Q_i}{\bar{P}_{i+1}}.$$

A routine computation shows that the equality holds for $\rho_j(\sigma)$ and any $\sigma \in [\sigma_j, \sigma_{j-1}]$ with $j = 2, \ldots, n$. Therefore, for any fixed σ, the inequality holds strictly for any other ρ. The converse, that is, $\phi(\rho, \sigma) = 1$ implies $(\rho, \sigma) = (\rho_j(\sigma), \sigma)$, is immediate from the strict monotonicity of ϕ. The proof of statement (iii) proceeds along the same lines of that of (ii) and is omitted. ■

Our previous result tells us how perturbation operates. According to Proposition 5, the optimal strategies perturb the user profile in such a manner that, in those categories with the lowest and highest ratios $\frac{q_k}{p_k}$, the apparent profile becomes proportional to the population's distribution. More precisely, the common ratio $\frac{t_k^*}{p_k}$ increases with both ρ and σ in those categories affected by forgery, that is, $k = 1, \ldots, i$. Exactly the opposite happens in those categories affected by suppression, where the common ratio $\frac{t_j^*}{p_j}$ decreases with both rates. This tendency continues until $\rho = \rho_{\text{crit}}(\sigma)$, at which point $t^* = p$. Fig. 3 illustrates this proportionality property in the case of the example depicted in Fig. 2.

C. Critical-Privacy Region

One of the results of Theorem 3 is that the boundary of the critical-privacy region is determined by the critical forgery-suppression threshold $\rho_j(\sigma)$, which we also denote by $\rho_{\text{crit}}(\sigma)$ to highlight this fact. The following proposition leverages on this result and characterizes said region. In particular, Proposition 6 first examines some properties of this threshold and then investigates the convexity of the critical-privacy region.

Proposition 6 (Convexity of the Critical-Privacy Region):

(i) ρ_j is a convex, piecewise linear function of $\sigma \in [\sigma_j, \sigma_{j-1}]$ for $j = 2, \ldots, n$.

(ii) $\mathscr{C}$ is convex.

Proof: From Theorem 3, it is routine to check the continuity of ρ_j on $[\sigma_n, \sigma_1]$. To show its convexity, we conveniently write this function as $\rho_j(\sigma) = m_j \sigma + b_j$, where $m_j = -\frac{P_{j-1}}{\bar{P}_j}$ and $b_j = \frac{P_{j-1} - Q_{j-1}}{\bar{P}_j}$. Next, we prove that the slopes satisfy $m_j < m_{j-1}$ for all $j = 3, \ldots, n$. We proceed by contradiction, assuming that $m_j \geqslant m_{j-1}$. Note that this inequality is equivalent to $P_{j-1} \bar{P}_{j-1} \leqslant \bar{P}_j - \bar{P}_j \bar{P}_{j-1}$ and, after algebraic simplification, to $p_{j-1} \leqslant 0$. This contradicts the positivity assumption (2), which, in turn, implies that $m_j < 0$ for all $j = 2, \ldots, n$. Therefore, since ρ_j is a piecewise linear function defined by the strictly increasing sequence of negative slopes $\{m_n, \ldots, m_2\}$, we can conclude that ρ_j is convex. This proves statement (i). The second statement follows from the first one. As ρ_j is convex, so is its epigraph, i.e., the critical-privacy region. ■

The conclusions drawn from Proposition 6 are illustrated in Fig. 4. In this figure we represent the critical and noncritical-privacy regions for $n = 5$ categories of interest; the distributions q and p assumed in this conceptual example are different

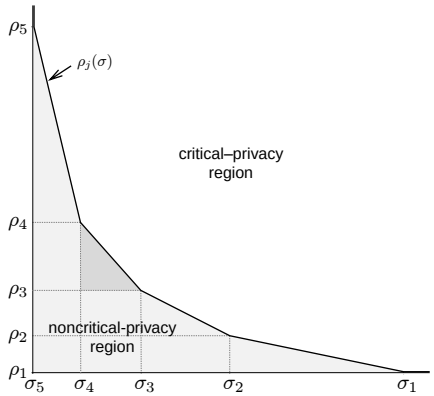


Fig. 4: Conceptual plot of the critical and noncritical privacy regions for $n = 5$ categories.

from those considered in Figs. 2 and 3. That said, the figure in question shows a straightforward consequence of our previous proposition—the noncritical-privacy region is nonconvex.

In this illustrative example, the sequences of forgery thresholds $\{\rho_1 \ldots, \rho_5\}$ and suppression thresholds $\{\sigma_5, \ldots, \sigma_1\}$ are strictly increasing. By Proposition 2, we can conclude then that the inequalities of the labeling assumption (3) hold strictly. Related to these thresholds is also the number of nonzero components of the optimal strategies, as follows from Theorem 3. Fig. 4 shows the sets of pairs (ρ, σ) where the number of nonzero components of r^* and s^* is fixed. Thus, in the triangular area shown darker, corresponding to the Cartesian product of the intervals $[\rho_3, \rho_4]$ and $[\sigma_4, \sigma_3]$, the solutions r^* and s^* have $i = 3$ and $n - j + 1 = 2$ nonzero components, respectively.

D. Case of Low Forgery and Suppression

This subsection characterizes the privacy-forgery-suppression function in the special case when $\rho, \sigma \simeq 0$.

Proposition 7 (Low Rates of Forgery and Suppression): Assume the nontrivial case in which $q \neq p$. Then, there exist two indexes i, j such that $0 = \rho_1 = \cdots = \rho_i < \rho_{i+1}$ and $0 = \sigma_n = \cdots = \sigma_j < \sigma_{j-1}$. For any $\rho \in [0, \rho_{i+1}]$ and $\sigma \in [0, \sigma_{j-1}]$, the number of nonzero components of the optimal forgery and suppression strategies is i and $n - j + 1$, respectively. Further, the gradient of the privacy-forgery-suppression function at the origin is

$$\nabla \mathcal{R}(0,0) = \begin{pmatrix} \frac{\partial \mathcal{R}(0,0)}{\partial \rho} \\ \frac{\partial \mathcal{R}(0,0)}{\partial \sigma} \end{pmatrix} = \begin{pmatrix} \log \frac{q_1}{p_1} - \mathrm{D}(q \,\|\, p) \\ \mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n} \end{pmatrix}.$$

Proof: The existence of the indexes i and j is guaranteed by the assumption that $q \neq p$. The number of nonzero components of r^* and s^* is trivial from Theorem 3. In view of this theorem, for any $\rho \in [0, \rho_{i+1}]$ and $\sigma \in [0, \sigma_{j-1}]$, we have

$$\mathcal{R}(\rho, \sigma) = \mathrm{D}\left(\frac{\tilde{q} + \rho(1, 0, \ldots, 0) - \sigma(0, \ldots, 0, 1)}{1 + \rho - \sigma} \middle\| \tilde{p} \right).$$

The continuity of the components of r^* and s^* proven in Corollary 4 (ii) ensures the continuity of the privacy-forgery-suppression function on $\bar{\mathscr{C}}$. It is routine to check its differentiability in this region and to obtain its derivative with respect to σ at the origin,

$$\frac{\partial \mathcal{R}(0,0)}{\partial \sigma} = Q_i \log \frac{Q_i \, \bar{P}_j}{P_i \, \bar{Q}_j} + \sum_{k=i+1}^{j-1} q_k \log \frac{\bar{P}_j \, q_k}{\bar{Q}_j \, p_k}.$$

On account of Proposition 2, the conditions $\rho_1 = \cdots = \rho_i$ and $\sigma_j = \cdots = \sigma_n$ imply

$$\frac{q_1}{p_1} = \cdots = \frac{q_i}{p_i} = \frac{Q_i}{P_i}$$

and

$$\frac{q_j}{p_j} = \cdots = \frac{q_n}{p_n} = \frac{\bar{Q}_j}{\bar{P}_j}.$$

Therefore,

$$\begin{aligned} \frac{\partial \mathcal{R}(0,0)}{\partial \sigma} &= \sum_{k=1}^{j-1} q_k \log \frac{q_k}{p_k} - Q_{j-1} \log \frac{q_n}{p_n} \\ &= \mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n}. \end{aligned}$$

The derivative of $\mathcal{R}$ with respect to ρ at $\rho = \sigma = 0$ follows analogously. ■

Next, we shall derive an expression for the relative decrement of the privacy-risk function at $\rho, \sigma \simeq 0$. To this end, define the *forgery relative decrement factor*

$$\delta_\rho = -\frac{\frac{\partial \mathcal{R}(0,0)}{\partial \rho}}{\mathcal{R}(0,0)} = 1 - \frac{\log \frac{q_1}{p_1}}{\mathrm{D}(q \,\|\, p)},$$

and the *suppression relative decrement factor*

$$\delta_\sigma = -\frac{\frac{\partial \mathcal{R}(0,0)}{\partial \sigma}}{\mathcal{R}(0,0)} = \frac{\log \frac{q_n}{p_n}}{\mathrm{D}(q \,\|\, p)} - 1.$$

By dint of Proposition 7, the first-order Taylor approximation of function (1) around $\rho = \sigma = 0$ yields

$$\mathcal{R}(\rho, \sigma) \simeq \mathrm{D}(q \,\|\, p) + \rho \left(\log \frac{q_1}{p_1} - \mathrm{D}(q \,\|\, p) \right) + \sigma \left(\mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n} \right),$$

or more compactly, in terms of the decrement factors,

$$\frac{\mathrm{D}(q \,\|\, p) - \mathcal{R}(\rho, \sigma)}{\mathrm{D}(q \,\|\, p)} \simeq \delta_\rho \, \rho + \delta_\sigma \, \sigma.$$

In words, the minimum and maximum ratios $\frac{q_k}{p_k}$ characterize the relative reduction in privacy risk. The following result, Proposition 8, establishes a bound on these relative decrement factors.

Proposition 8 (Relative Decrement Factors): In the nontrivial case when $q \neq p$, the relative decrement factors satisfy $\delta_\rho > 1$ and $\delta_\sigma > 0$.

Proof: Observe that the statement $\delta_\rho > 1$ is equivalent to the condition $q_1 < p_1$. We prove this by contradiction. Suppose that $q_1 > p_1$. By the labeling assumption (3), it follows that $q_k > p_k$ for all k, what leads to the contradiction that $1 = \sum q_k > \sum p_k = 1$. Now assume that $q_1 = p_1$. Since $q \neq p$, there must exist an index i such that

$$\frac{q_1}{p_1} = \cdots = \frac{q_{i-1}}{p_{i-1}} < \frac{q_i}{p_i} \leqslant \cdots \leqslant \frac{q_n}{p_n}.$$

But this implies that

$$1 - \sum_{k=1}^{i-1} q_k = \sum_{k=i}^{n} q_k > \sum_{k=i}^{n} p_k = 1 - \sum_{k=1}^{i-1} q_k,$$

a contradiction. This proves the first part of the proposition.

For the second part, note that the statement $\delta_\sigma > 0$ is equivalent to

$$q_1 \log \frac{q_1}{p_1} + \cdots + q_n \log \frac{q_n}{p_n} < \log \frac{q_n}{p_n},$$

and, after algebraic manipulation, to

$$q_1 \log \frac{q_1}{p_1} \frac{p_n}{q_n} + \cdots + q_{n-1} \log \frac{q_{n-1}}{p_{n-1}} \frac{p_n}{q_n} < 0.$$

The positivity and labeling assumptions (2), (3) ensure that all terms in the sum are nonpositive. However, the additional assumption $q \neq p$ implies that $\frac{q_1}{p_1} < \frac{q_n}{p_n}$, which in turn implies that the first term is negative and so is, consequently, the entire summation. ■

Conceptually, the bound on δ_ρ tells us that the relative decrement in privacy risk is greater than the forgery rate introduced. This is under the assumption that $q \neq p$ and at low rates of forgery and suppression. The bound on δ_σ, however, is looser than the previous one and just ensures that an increase in the suppression rate always leads to a decrease in privacy risk, as one would expect.

E. Pure Strategies

In the previous subsections we investigated the forgery and the suppression of ratings as a *mixed* strategy that users may adopt to enhance their privacy. In this subsection we contemplate the case in which users may be reluctant to use these two mechanisms in conjunction; and as a consequence, they may opt for a *pure* strategy consisting in the application of either forgery or suppression. In this case, it would be useful to determine which is the most appropriate technique in terms of the privacy-utility trade-off posed. Our next result, Corollary 9, provides some insight on this, under the assumption that, from the user's perspective, the impact on utility due to forgery is equivalent to that caused by the effect of suppression.

Before showing this result, observe from Theorem 3 that $\rho_n = \frac{q_n}{p_n} - 1$ is the minimum forgery rate such that $\mathcal{R}(\rho, 0) = 0$. Analogously, $\sigma_1 = 1 - \frac{q_1}{p_1}$ is the minimum suppression rate satisfying $\mathcal{R}(0, \sigma) = 0$. In other words, ρ_n and σ_1 are the *critical rates* of the pure forgery and suppression strategies, respectively. Further, note that $\sigma_1 < \sigma_0 = 1$, on account of the positivity assumption (2). However, $\rho_n > 1$ if, and only if, $\frac{q_n}{p_n} > 2$.

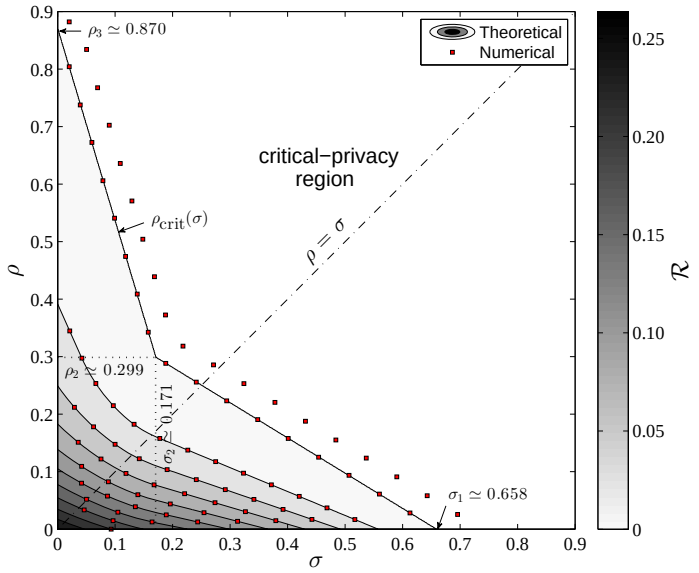


Fig. 5: Contour lines of the privacy-forgery-suppression function, the corresponding forgery and suppression thresholds, and the critical and noncritical privacy regions.

Corollary 9 (Pure Strategies): Consider the nontrivial case when $q \neq p$.

(i) The critical rates of the pure forgery and suppression strategies satisfy $\rho_n < \sigma_1$ if, and only if,

$$\frac{q_1/p_1 + q_n/p_n}{2} < 1.$$

(ii) The forgery and the suppression relative decrement factors satisfy $\delta_\rho > \delta_\sigma$ if, and only if,

$$\sqrt{\frac{q_1}{p_1}\,\frac{q_n}{p_n}} < 2^{\mathrm{D}(q\,\|\,p)}.$$

Proof: Both statements are immediate from the definitions of ρ_n and σ_1 on the one hand, and δ_ρ and δ_σ on the other. ■

In conceptual terms, the condition $\rho_n < \sigma_1$ means that the pure forgery strategy is the most appropriate mechanism in terms of causing the minimum distortion to attain the critical-privacy region. On the other hand, the condition $\delta_\rho > \delta_\sigma$ implies that, at low rates, the pure forgery strategy offers better privacy protection than the pure suppression strategy does. Therefore, the conclusion that follows from Corollary 9 is that, together with the quantity $\mathrm{D}(q\,\|\,p)$, the arithmetic and geometric mean of the ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$ determine which strategy to choose.

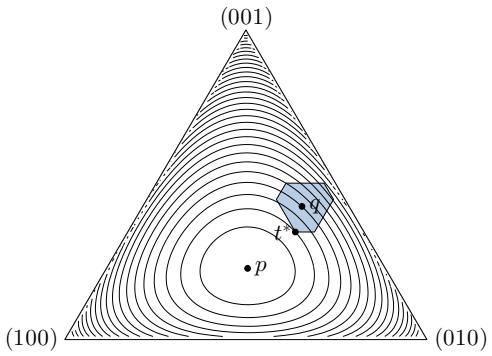
Another interesting remark is the duality of these two ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$. The former characterizes the minimum rate for the pure suppression strategy to reach the critical-privacy region and, at the same time, it establishes the privacy gain at low forgery rates. Conversely, the latter ratio defines the critical rate of the pure forgery strategy and determines the relative decrement in privacy risk at low suppression rates.

Lastly, we would like to establish a connection between our work and that of [11], [20], where the *pure* forgery and suppression strategies are investigated. Denote by $\mathcal{R}_{\mathrm{F}}$ the function derived in [11] modeling the trade-off between forgery rate and privacy *risk*, the latter being measured as the KL divergence between the user's apparent profile and the population's distribution. Define ρ' as the ratio of forged ratings to *total* number of ratings. Accordingly, it can be shown that $\rho' = \frac{\rho}{1+\rho}$ and that $\mathcal{R}(\rho, 0) = \mathcal{R}_{\mathrm{F}}(\rho')$. On the other hand, denote by $\mathcal{P}_{\mathrm{S}}$ the function in [20] characterizing the trade-off between suppression rate and privacy *gain*. In this case, privacy is measured as the Shannon's entropy of the user's apparent profile. Under the assumption that the population's profile is uniform, it can be proven that $\mathcal{R}(0, \sigma) = \log n - \mathcal{P}_{\mathrm{S}}(\sigma)$. In short, our formulation of the problem of optimal forgery and suppression of ratings encompasses, as particular cases, the cited works.

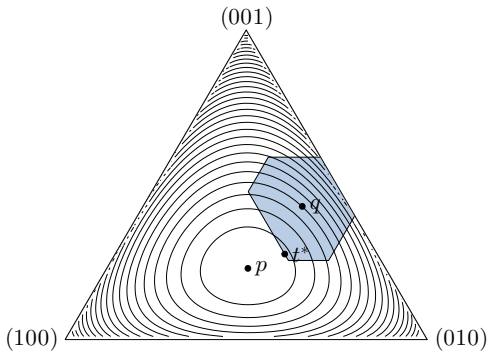
F. Numerical Example

This subsection presents a numerical example that illustrates the theoretical analysis conducted in the previous subsections. Later on in Sec. V we shall evaluate the effectiveness of our approach in a real scenario, namely in the movie recommendation system *Movielens*. In our numerical example we assume $n = 3$ categories of interests. Although the example shown here is synthetic, these three categories could very well represent interests across topics such as technology, sports and beauty. Accordingly, we suppose that the user's rating distribution is

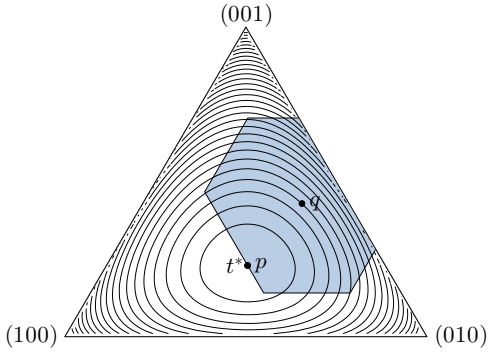
$$q = (0.130, 0.440, 0.430),$$



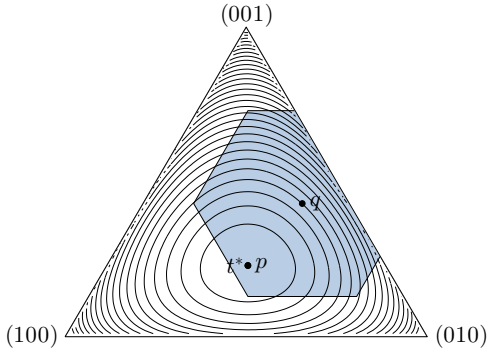
(a) $\rho = 0.050$, $\sigma = 0.100$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.093$, $\mathcal{R}(\rho,\sigma) \simeq 0.131$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.498$, $r^* = (0.050, 0, 0)$, $s^* = (0, 0, 0.100)$, $t^* \simeq (0.189, 0.463, 0.347)$.



(b) $\rho = 0.100$, $\sigma = 0.200$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.356$, $\mathcal{R}(\rho,\sigma) \simeq 0.050$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.190$, $r^* = (0.100, 0, 0)$, $s^* \simeq (0, 0.019, 0.181)$, $t^* \simeq (0.256, 0.468, 0.276)$.



(c) $\rho \simeq 0.219$, $\sigma = 0.300$, $\rho/\rho_{\text{crit}}(\sigma) = 1$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$, $r^* \simeq (0.219, 0, 0)$, $s^* \simeq (0, 0.081, 0.219)$, $t^* = p$.



(d) $\rho = 0.300$, $\sigma = 0.300$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 1.368$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$, $r^* \simeq (0.260, 0.021, 0.019)$, $s^* = (0.010, 0.071, 0.219)$, $t^* = p$.

Fig. 6: Probability simplices showing, for several interesting values of ρ and σ, the user's actual profile $q = (0.130, 0.440, 0.430)$, the population's distribution $p = (0.380, 0.390, 0.230)$, the optimal apparent distribution t^* and the set of feasible apparent distributions.

and the population's,

$$p = (0.380, 0.390, 0.230).$$

Note that these distributions satisfy the positivity and labeling assumptions (2), (3).

From Sec. IV-A, we easily obtain the forgery thresholds $\rho_1 = 0$, $\rho_2 \simeq 0.299$ and $\rho_3 \simeq 0.870$ on the one hand, and on the other the suppression thresholds $\sigma_3 = 0$, $\sigma_2 \simeq 0.171$ and $\sigma_1 \simeq 0.658$. The thresholds ρ_3 and σ_1 are the critical rates of the pure strategies. If we are to reach the critical-privacy region and do not have any preference for either forgery or suppression, the fact that $\rho_3 > \sigma_1$ leads us to opt for suppression as pure strategy. However, the geometric mean of $\frac{q_1}{p_1}$ and $\frac{q_3}{p_3}$ is approximately 0.799, which is lower than $2^{\mathrm{D}(q\,\|\,p)} \simeq 1.20$. On account of Corollary 9, this means that the pure forgery strategy contributes to a greater reduction in privacy risk at low rates than suppression does. In fact, the gradient of the privacy-forgery-suppression function at the origin is $\nabla\mathcal{R}(0,0)^{\mathrm{T}} \simeq (-1.81, -0.639)$, by virtue of Proposition 7.

Fig. 5 shows the contour lines of this function, computed analytically from Theorem 3 and numerically [(b)]. The region plotted in gray shades corresponds to the noncritical-privacy region $\bar{\mathscr{C}}$. The initial privacy risk is $\mathcal{R}(0,0) \simeq 0.263$. The white area represents the critical-privacy region $\mathscr{C}$, where the apparent user profile coincides with the population's distribution and thus the privacy risk vanishes. An interesting observation arising from Fig. 5 is the synergistic effect of combining forgery and suppression. Just as an example, in the case when $\rho = \rho_2$ and $\sigma = \sigma_2$, the sum of these two distortion measures is lower than the critical rates of the pure strategies.

Next, we examine the optimal apparent rating distribution for different values of ρ and σ. For this purpose, the user's genuine distribution q, the population's distribution p and the optimal apparent distribution t^* are depicted in the probability simplices shown in Fig. 6. In each simplex, we also represent the contour lines of the KL divergence $\mathrm{D}(\cdot\,\|\,p)$ between every distribution in the simplex and p. Further, we plot the set of feasible apparent user distributions, not necessarily optimal, for four different combinations of ρ and σ; in any of these cases, the set takes the form of a hexagon. Having said this, now we turn our attention to Fig. 6(a). In this case, the optimal forgery and suppression strategies have $i = n - j + 1 = 1$ nonzero component, since $\rho \in [0, \rho_2]$ and $\sigma \in [0, \sigma_2]$. This places the solution t^* at one vertex of the hexagon. A remarkable fact is that, for these rates,

[(b)]The numerical method chosen is the interior-point algorithm [41] implemented by the Matlab R2012b function `fmincon`.

TABLE I: Category index of the particular user examined in our experiments. The categories of *Movielens* have been sorted and indexed in order to satisfy the labeling assumption (3).

Index	Category name	Index	Category name	Index	Category name
1	animation	7	sci-fi	13	war
2	action	8	comedy	14	mystery
3	film-noir	9	thriller	15	musical
4	children's	10	fantasy	16	romance
5	adventure	11	horror	17	IMAX
6	crime	12	western	18	drama
				19	documentary

the privacy risk is approximately halved. In the end, consistently with Proposition 8, the forgery and the suppression relative decrement factors are $\delta_\rho \simeq 6.87 > 1$ and $\delta_\sigma \simeq 2.42 > 0$.

In the case shown in Fig. 6(b), r^* still has $i = 1$ nonzero components, while s^* contains $n - j + 1 = 2$ nonzero components. Geometrically, the optimal apparent distribution lies at one edge of the feasible region. This lowers privacy risk to a 19% of its initial value. The case in which $(\rho, \sigma) = (\rho_{\text{crit}}(\sigma), \sigma)$ is depicted in Fig. 6(c). Here, the number of nonzero components of r^* and s^* remains the same as in the previous case, but the privacy risk becomes zero. The last case, illustrated in Fig. 6(d), does not have any practical application, as $\mathcal{R}(\rho, \sigma) = 0$ for any $(\rho, \sigma) \in \partial\mathscr{C}$. In this figure we can observe that the solution t^* is placed in the interior of the hexagon, and that the orthogonality principle of the strategies r^* and s^* stated in Corollary 4 is not satisfied.

V. Experimental Evaluation

In this section we evaluate the extent to which the forgery and the suppression of ratings could enhance user privacy in a real-world recommendation system. The system chosen to conduct this evaluation is *Movielens*, a popular movie recommender developed by the GroupLens Research Lab [42] at the University of Minnesota. As many other recommenders, *Movielens* allows users to both rate and tag movies according to their preferences. These preferences are then exploited by the recommender to suggest movies that users have not watched yet.

A. Data set

The data set that we used to assess our data-perturbative mechanism is the *Movielens 10M* data set [43], which contains 10 000 054 ratings and 95 580 tags. The ratings and tags included in this data set were assigned to 10 681 movies by 71 567 users. The data are organized in the form of quadruples (*username*, *movie*, *rating*, *time*), each one representing the action of a user rating a movie at a certain time. Usernames have been replaced with numbers in an attempt to anonymize the data set.

For our purposes of experimentation, we just needed the data fields *username* and *movie*, together with the categories each movie belongs to. *Movielens* contemplates $n = 19$ categories or movies genres, listed in alphabetical order as follows: *action*, *adventure*, *animation*, *children's*, *comedy*, *crime*, *documentary*, *drama*, *fantasy*, *film-noir*, *horror*, *IMAX*, *musical*, *mystery*, *romance*, *sci-fi*, *thriller*, *war* and *western*. As we shall see later in Sec. V-B, for each particular user, we shall have to rearrange those categories in such a way that the labeling assumption (3) is satisfied.

In our data set, all users rated, at least, 20 movies. This was the minimum number of ratings for the recommender to start working [(c)]. After the elimination of those users who exclusively tagged movies, the total number of users reduced to 69 878. Despite the large number of users, we found that only 4 099 satisfied the positivity assumption (2). Considering that this small group of users represents just the 5.8% of the total number of users, we can assume that the application of our technique will have a negligible effect on the population's profile p, as supposed in Sec. III-D.

B. Results

In this subsection we examine how the forgery and the suppression of ratings may help users of *Movielens* to enhance their privacy. With this aim, first, we analyze the effect of the perturbation of ratings on the privacy protection of a particular user from our data set. Secondly, we consider the entire set of 4 099 users and assess the relative reduction in privacy risk when these users apply the same forgery and suppression rates. Lastly, we investigate the forgery and the suppression strategies separately, and draw some conclusions about these two pure strategies.

To conduct our first experiments, we choose a particular user from our data set [(d)]. Before perturbing the movie rating history of this user, it is necessary that the components of the user's profile q and the population's distribution p be rearranged to satisfy the labeling assumption (3). Table I shows how movie categories have been sorted, and then indexed from 1 to n, to fulfill the assumption above. We would like to note that the index provided in this table does not have to coincide with the index of other users in our data set.

[(c)] Nowadays, the algorithm implemented by *Movielens* requires only 15 ratings to start generating predictions.

[(d)] The user considered in this first series of experiments is identified by the number `3301` in [43].

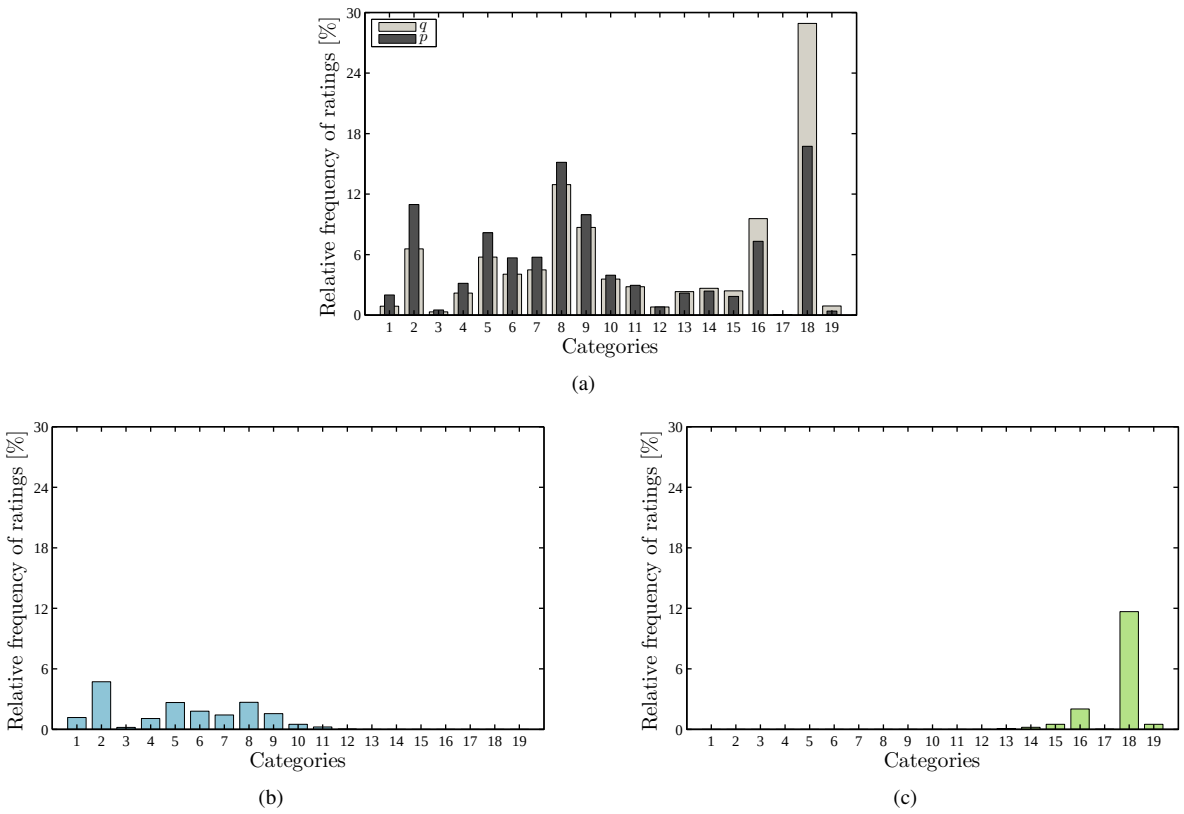


Fig. 7: In this figure we represent (a) the item distribution q of a particular user as well as the population's item distribution p. In addition, we plot (b) the optimal forgery strategy r^* and (c) the optimal suppression strategy s^* that the user in question should adopt when they specify $\sigma = 0.150$ and $\rho = \rho_{\text{crit}}(\sigma) \simeq 0.180$.

Fig. 7(a) depicts the user profile and the population profile, the latter being computed by averaging across the 69 878 users. From this figure we note that the user's interest far exceeds the population's in categories such as *musical*, *romance*, *IMAX*, *drama* and *documentary*. More precisely, such ratios $\frac{q_k}{p_k}$ yield

$$\left(\frac{q_k}{p_k}\right)_{k=15,\ldots,19} \simeq (1.300, 1.306, 1.451, 1.728, 2.292).$$

In this figure, we also observe that the user's interest and the population's in the category 17 are nearly zero, namely $q_{17} \simeq 0.0005$ and $p_{17} \simeq 0.0003$.

On the other hand, Fig. 7(a) indicates that the user shows little interest, compared to the population's preferences, in categories such as *animation*, *action*, *film-noir* or *children's*, to name just a few. Specifically, the first six smallest ratios $\frac{q_k}{p_k}$ yield

$$\left(\frac{q_k}{p_k}\right)_{k=1,\ldots,6} \simeq (0.444, 0.599, 0.651, 0.691, 0.705, 0.714).$$

Figs. 7(b) and 7(c) show the optimal forgery and suppression strategies that this particular user should apply, in the case when $\sigma = 0.150$ and $\rho_{\text{crit}}(\sigma) \simeq 0.180$. The solutions plotted in these figures are consistent with our two previous observations—the optimal forgery strategy recommends that the user submit false ratings to movies falling into the categories where the ratio $\frac{q_k}{p_k}$ is low; and the optimal suppression strategy suggests that the user refrain from rating movies belonging to categories where the ratio $\frac{q_k}{p_k}$ is high. Just as an example, the fact that $s^*_{17} \simeq 0.0001$ means that the user at hand should eliminate one in five ratings to movies classified as *IMAX*.

The optimal trade-off surface among privacy, forgery rate and suppression rate is represented in Fig. 8. In this figure we plot the contour levels of the function $\mathcal{R}(\rho, \sigma)$, which we computed theoretically. The initial privacy risk is $\mathcal{R}(0, 0) \simeq 0.101$ and the arithmetic mean between the ratios $\frac{q_1}{p_1}$ and $\frac{q_{19}}{p_{19}}$ yields approximately 1.37. Since the mean is higher than 1, Corollary 9 tells us that the user should opt for suppression as pure strategy, in lieu of forgery. This is under the assumption that they wish to achieve the minimum privacy risk and do not have any preference for any of the pure strategies. Nevertheless, the fact that $\delta_\rho \simeq 12.6 > \delta_\sigma \simeq 10.9$ leads us to choose forgery as pure strategy for $\rho, \sigma \simeq 0$. When both strategies are combined, note that a forgery and suppression rate of just 0.1% leads to a relative reduction in privacy risk of 2.35%, on account of the first-order Taylor approximation derived in Sec. IV-D.

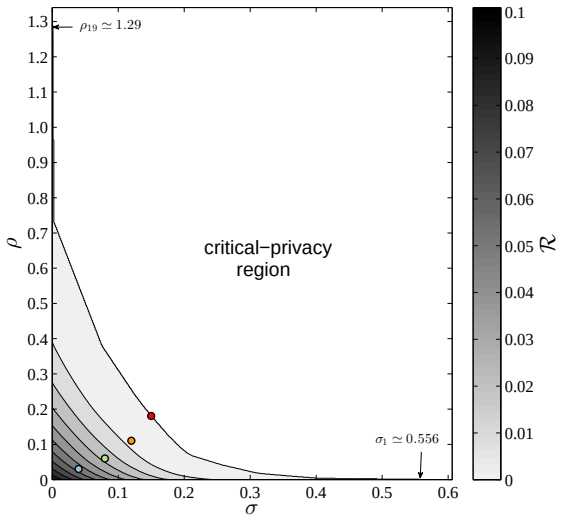


Fig. 8: Optimal trade-off surface among privacy, forgery rate and suppression rate for one particular user in our data set. The four points shown in this figure correspond to the pairs of values (ρ, σ) that we used to show the proportionality relationship between t^* and p in Fig. 9.

In Fig. 8 we have also plotted 4 points, which correspond to the following pairs of values (ρ, σ): $(0.03, 0.04)$, $(0.06, 0.08)$, $(0.11, 0.12)$ and $(0.18, 0.15)$. For each of these pairs, we have represented the quotient $\frac{t_k^*}{p_k}$ in Fig. 9. The aim is to show how the optimal apparent profile becomes proportional to the population's distribution, as the user approaches the critical-privacy region. Fig. 9(a) considers the first pair of values. Here, ρ and σ fall into the intervals $[\rho_6, \rho_7]$ and $[\sigma_{18}, \sigma_{17}]$, respectively. Consistently with Proposition 5, we check that $\frac{t_1^*}{p_1} = \cdots = \frac{t_6^*}{p_6} \simeq 0.756$ and that $\frac{t_{18}^*}{p_{18}} = \frac{t_{19}^*}{p_{19}} \simeq 1.52$.

In Fig. 9(b) we double the rates of forgery and suppression. On the one hand, this leads to $\frac{t_1^*}{p_1} = \cdots = \frac{t_7^*}{p_7}$. On the other, the fact that $\sigma \in [\sigma_{15}, \sigma_{14}]$ implies that $\frac{t_{15}^*}{p_{15}} = \cdots = \frac{t_{19}^*}{p_{19}}$. It is also interesting to note that, for these relatively small values of ρ and σ, the final privacy risk is 26% of the initial value $\mathrm{D}(q \,\|\, p)$.

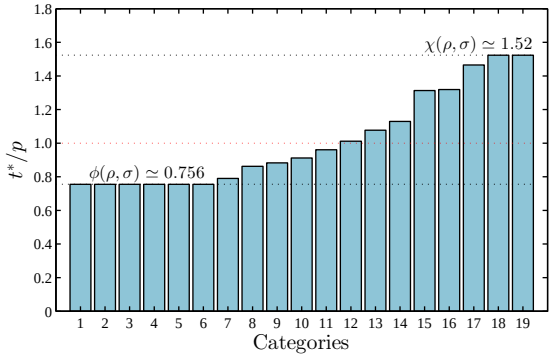
As ρ and σ increase, so does the function ϕ. The contrary happens with the function χ, which decreases with both rates. In Fig. 9(c), for example, the proportionality relationship between t^* and p holds for all except 4 categories. The last pair $(\rho, \sigma) \simeq (0.18, 0.15)$ lies at the boundary of $\mathscr{C}$, as shown in Fig. 8. This implies that $\frac{t^*}{p} = 1$ and therefore that $\mathcal{R}(\rho, \sigma) = 0$, as captured in Fig. 9(d).

Having examined the case of a specific user, in our next series of experiments we evaluate the privacy-protection level that users can achieve if they are disposed to forge and eliminate a fraction of their ratings. For simplicity, we suppose that all users satisfying the positivity assumption (2) apply a common forgery rate and a common suppression rate. Fig. 10 depicts the contours of the 10th, 50th and 90th percentile surfaces of relative reduction in privacy risk, for different values of ρ and σ. Two conclusions can be drawn from this figure.

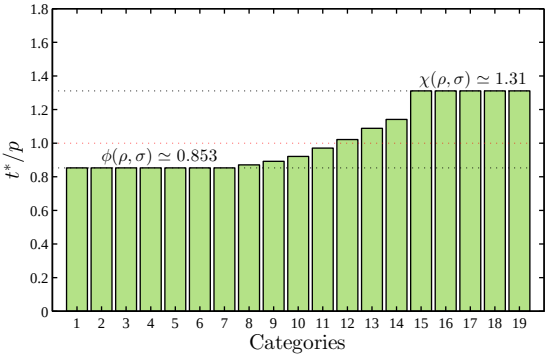
- First, for relatively small values of ρ and σ (lower than 15%), a vast majority of users lowered privacy risk significantly. In quantitative terms, we observe in Fig. 10(a) that, for $\rho = \sigma = 0.05$, the 10% of users adhered to our technique obtained a reduction in privacy risk by at least 52.4%. For those same rates of forgery and suppression rates, the 50th and 90th percentiles are 73.9% and 94.8%. For higher rates, e.g., $\rho = \sigma = 0.15$, Fig. 10(b) highlights that half of users experienced a reduction in privacy risk less than or equal to 100%.
- Secondly, the three percentile surfaces exhibit a certain symmetry with respect to the line $\rho = \sigma$. If this symmetry were exact, the exchange of the rates of forgery and suppression would not have any impact on the resulting privacy-protection achieved. However, this is not the case. For example, Fig. 10(a) shows a lower reduction in privacy risk for $\rho < \sigma$, particularly accentuated when $\sigma \simeq 0$. The reason for this may be found in the fact that, for most users, ρ_n is greater than σ_1. We shall elaborate more on this later on when we consider forgery and suppression as pure strategies.

Next, we analyze the privacy protection provided by our technique for $\rho, \sigma \simeq 0$. In the theoretical analysis conducted in Sec. IV-D we derived an expression for the relative reduction in privacy risk at low rates. Particularly, said expression was in terms of two factors, namely δ_ρ and δ_σ. In Fig. 11 we show the probability distribution of these factors. Consistently with Proposition 8, the minimum values of these factors are $\delta_\rho \simeq 3.12 > 1$ and $\delta_\sigma \simeq 2.30 > 0$. The maximum values attained by these forgery and the suppression factors are approximately 324.98 and 266.13. On the other hand, in favour of suppression is the fact that the percentage of users with $\delta_\rho \geqslant 30$ is lower than those users with $\delta_\sigma \geqslant 30$. More precisely, these percentages yield 26.8% and 33.1%, respectively. In the end, an eye-opening finding is that $\delta_\rho > \delta_\sigma$ in 43.45% of users, which suggests introducing a suppression rate higher than that of forgery, at least at low rates.

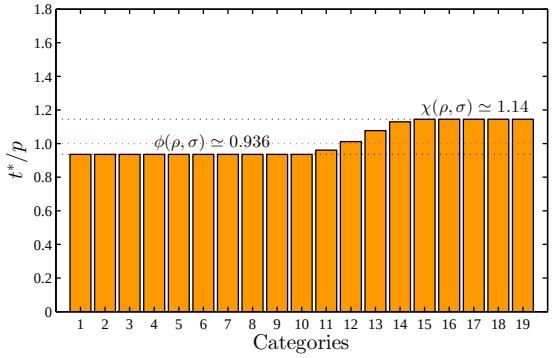
After analyzing the forgery and the suppression of ratings as a mixed strategy, our last experimental results contemplate the application of forgery and suppression as pure strategies. In Fig. 12 we illustrate the probability distribution of the critical rates ρ_n and σ_1. The critical forgery rate ranges approximately from 0.171 to 54.18, and its average is 3.45. The critical suppression



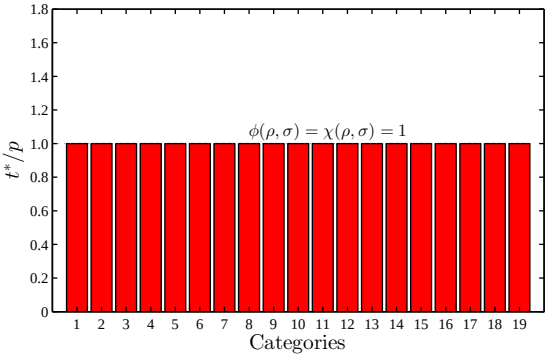
(a) $\rho = 0.03$, $\rho \in [\rho_6, \rho_7]$, $\sigma = 0.04$, $\sigma \in [\sigma_{18}, \sigma_{17}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.055$, $\mathcal{R}(\rho,\sigma) \simeq 0.055$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.549$.



(b) $\rho = 0.06$, $\rho \in [\rho_7, \rho_8]$, $\sigma = 0.08$, $\sigma \in [\sigma_{15}, \sigma_{14}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.164$, $\mathcal{R}(\rho,\sigma) \simeq 0.026$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.259$.



(c) $\rho = 0.11$, $\rho \in [\rho_{10}, \rho_{11}]$, $\sigma = 0.12$, $\sigma \in [\sigma_{15}, \sigma_{14}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.434$, $\mathcal{R}(\rho,\sigma) \simeq 0.006$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.061$.



(d) $\rho \simeq 0.180$, $\rho \in [\rho_{12}, \rho_{13}]$, $\sigma = 0.15$, $\sigma \in [\sigma_{13}, \sigma_{12}]$, $\rho/\rho_{\text{crit}}(\sigma) = 1$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$.

Fig. 9: Proportionality relationship between, on the one hand, the optimal apparent item distribution t^* of the user identified as `3301` in our data set, and on the other, the population's item distribution p.

rate, on the other hand, goes from 0.153 to 0.963, and its average is 0.632. These figures indicate that, on average, a user will have either to refrain from rating an item six out of ten times, or submit nearly 3.45 false ratings per each original rating. This is, of course, when the user wishes to reach the critical-privacy region. Bearing these figures in mind, it is not surprising then that 95.3% of the users in our data set would opt for suppression as pure strategy, as it comes at the cost of a lower impact on utility.

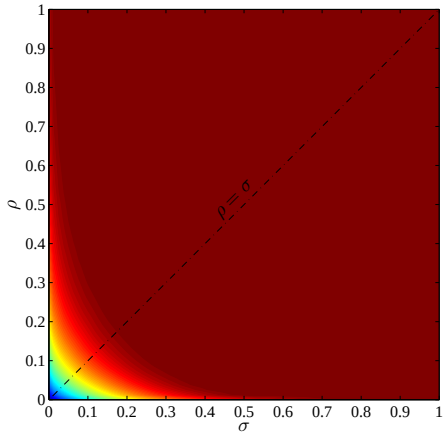
VI. Conclusion

In the literature of recommendation systems there exists a variety of approaches aimed at protecting user privacy. Among these approaches, the forgery and the suppression of ratings emerge as a technique that may hinder attackers in their efforts to accurately profile users on the basis of the items they rate. Our technique does not require that users trust neither the recommender nor the network operator, it is simple in terms of infrastructure requirements, and it can be used in combination with other approaches providing soft privacy. However, as any data-perturbative approach, our privacy-enhancing technology comes at the expense of a loss in data utility, in particular a degradation of the quality of the recommender's predictions. Put another way, it poses a trade-off between privacy and utility.

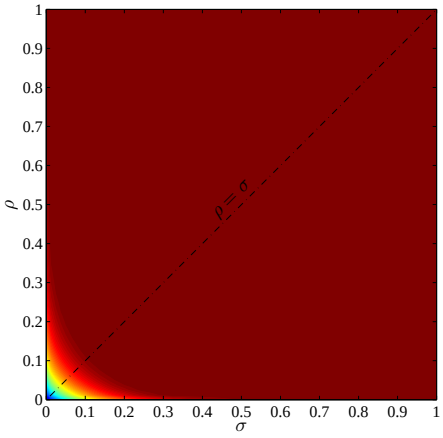
The objective of this paper is to investigate mathematically said trade-off. For this purpose, first we propose a quantitative measure of both privacy and utility. We quantify privacy risk as the KL divergence between the user's rating distribution and the population's, and measure utility as the fraction of ratings the user is willing to forge and suppress. With these two quantities, we formulate a multiobjective optimization problem characterizing the trade-off between privacy risk on the one hand, and on the other forgery rate and suppression rate.

Our theoretical analysis provides a closed-form solution to this problem and characterizes the optimal trade-off surface between privacy and utility. The solution is confined to the closure of the noncritical-privacy region. The interior of the critical-privacy region is of no interest as the privacy risk attains its minimum value at the boundary of $\mathscr{C}$. In the region of interest, our analysis finds that the optimal forgery and suppression strategies are orthogonal. In addition, these two strategies follow an intuitive principle. The forgery strategy recommends adding ratings to those categories where the user's interest is lower than the population's. The suppression strategy suggests eliminating those ratings belonging to the categories where the user shows too much interest compared to the reference distribution.

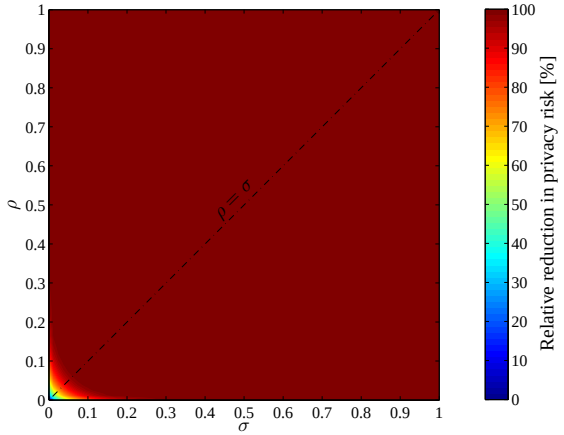
Our theoretical study also examines how these optimal strategies perturb user profiles. It is interesting to observe that the optimal apparent profile becomes proportional to the population's distribution in those categories with the lowest and highest



(a) 10th percentile.



(b) 50th percentile.



(c) 90th percentile.

Fig. 10: We assume that the 4 099 users satisfying the positivity assumption (2) protect their privacy by using a common forgery rate and a common suppression rate. Under this assumption, we plot some percentiles surfaces of relative reduction in privacy risk, against these two common rates.

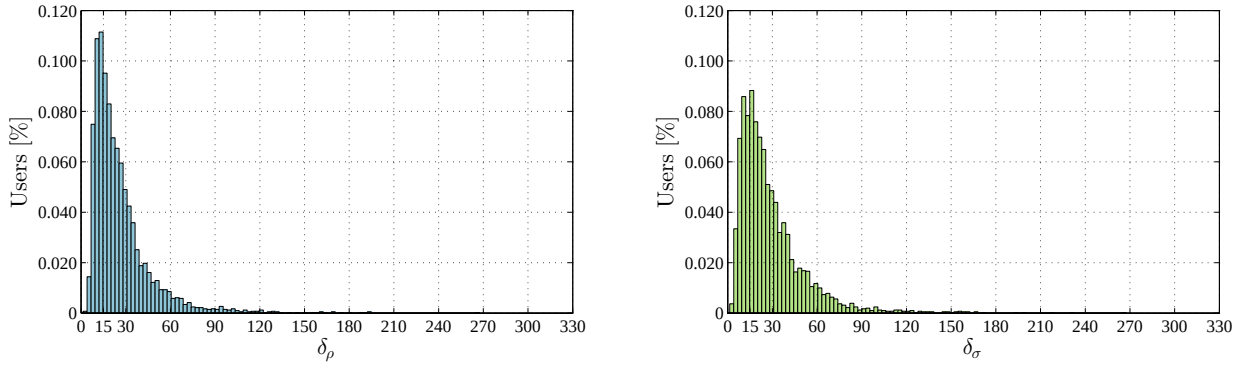


Fig. 11: Probability distribution of the relative decrement factors of forgery and suppression.

ratios $\frac{q_k}{p_k}$. Our analysis also includes the characterization of $\mathcal{R}$ at low rates of forgery and suppression. More accurately, we provide a first-order Taylor approximation of the privacy-utility trade-off function, from which we conclude that the ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$ determine, together with the quantity $\mathrm{D}(q \,\|\, p)$, the privacy risk at low rates. An eye-opening fact is that the relative decrement in privacy risk is greater than the forgery rate introduced.

Further, we consider the special case when forgery and suppression are not used in combination. Under this consideration, we investigate which one is the most appropriate technique, first, in terms of causing the minimum distortion to reach the critical-privacy region, and secondly, in terms of offering better privacy protection at low rates. Our findings show that the arithmetic and geometric mean of the maximum and minimum ratios $\frac{q_k}{p_k}$ play a fundamental role in deciding the best technique to use. Afterwards, our formulation and theoretical analysis are illustrated with a numerical example.

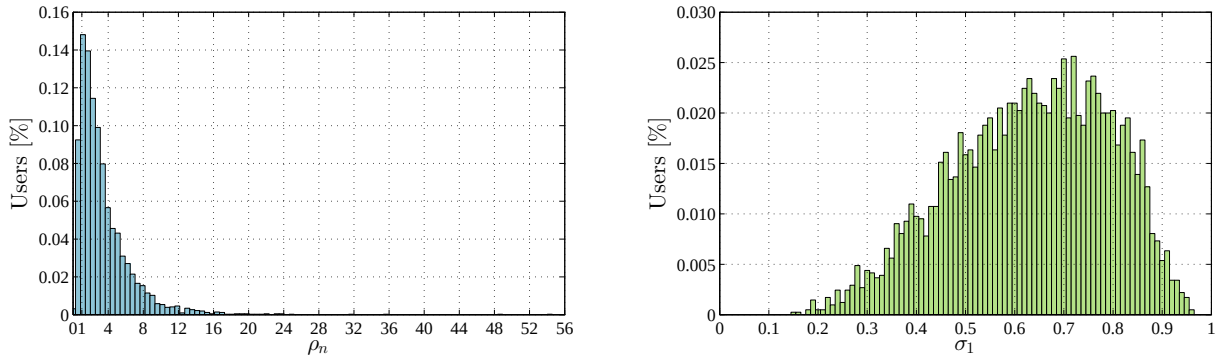


Fig. 12: Probability distribution of the critical forgery and suppression rates.

In the end, the last section is devoted to the experimental evaluation of our data-perturbative mechanism in a real-world recommendation system. In particular, we examine how the application of the forgery and the suppression of ratings may preserve user privacy in *Movielens*. Among other results, we find that a large majority of users significantly reduce privacy risk for forgery and suppression rates of just 15%. In our data set, the probability distributions of the relative decrement factors indicate that, at low rates, forgery provides a higher reduction in privacy risk than suppression does. By contrast, we observe that the suppression relative decrement factor is greater than that of forgery in 43.45% of users. Lastly, we consider the case when users must opt for either forgery or suppression; and find that the latter is the best strategy to use in 95.3% of users who wish to vanish privacy risk while causing the minimum distortion.

REFERENCES

[1] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-protecting architecture for collaborative filtering via forgery and suppression of ratings," in *Proc. Int. Workshop Data Priv. Manage., Auton. Spontaneus Secur. (DPM)*, Leuven, Belgium, Sep. 2011, pp. 42–57.

[2] U. Hanani, B. Shapira, and P. Shoval, "Information filtering: Overview of issues, research and systems," *User Model. User-Adap. Interact.*, vol. 11, no. 3, pp. 203–259, Aug. 2001.

[3] D. Oard and J. Kim, "Implicit feedback for recommender systems," in *Proc. AAAI Workshop Recommender Syst.*, 1998, pp. 81–83.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[5] L. F. Cranor, "'I didn't buy it for myself'. Privacy and e-commerce personalization," in *Proc. Workshop Priv. Electron. Soc.*, Washington, DC, 2003, pp. 111–117.

[6] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur., Priv. (SP)*. Washington, DC: IEEE Comput. Soc., 2008, pp. 111–125. [Online]. Available: http://dx.doi.org/10.1109/SP.2008.33

[7] "Netflix prize." [Online]. Available: http://en.wikipedia.org/wiki/Netflix_Prize

[8] J. Zaslow, "If TiVo thinks you are gay, here's how to set it straight," Nov. 2002. [Online]. Available: http://online.wsj.com/article_email/SB1038261936872356908.html

[9] S. Fox, "Trust and privacy online: Why americans want to rewrite the rules," Pew Internet, Amer. Life Project, Res. Rep., Aug. 2000.

[10] D. L. Hoffman, T. P. Novak, and M. Peralta, "Building consumer trust online," *Commun. ACM*, vol. 42, no. 4, pp. 80–85, Apr. 1999.

[11] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[12] D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "An information-theoretic privacy criterion for query forgery in information retrieval," in *Proc. Int. Conf. Secur. Technol.(SecTech)*, ser. Lecture Notes Comput. Sci. (LNCS). Jeju Island, South Korea: Springer-Verlag, Dec. 2011, pp. 146–154, invited paper.

[13] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Gen. Comput. Syst.*, 2013, to appear. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.001

[14] H. Polat and W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in *Proc. SIAM Int. Conf. Data Min. (SDM)*. IEEE Comput. Soc., 2003.

[15] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*. Washington, DC: IEEE Comput. Soc., 2003, pp. 99–106.

[16] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. ACM, 2005, pp. 37–48.

[17] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," in *Proc. ACM Int. Symp. Appl. Comput. (SASC)*. ACM, 2005, pp. 791–795.

[18] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Santa Barbara, CA, 2001, pp. 247–255.

[19] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-preserving architecture for the semantic web based on tag suppression," in *Proc. Int. Conf. Trust, Priv., Secur., Digit. Bus. (TRUSTBUS)*, Bilbao, Spain, Aug. 2010, pp. 58–68.

[20] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz, and O. Esparza, "Optimal tag suppression for privacy protection in the semantic web," *Data, Knowl. Eng.*, vol. 81–82, no. 0, pp. 46–66, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2012.07.004

[21] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forné, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Trans. Knowl. Data Eng.*, 2012, to appear. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2012.248

[22] J. Canny, "Collaborative filtering with privacy via factor analysis," in *Proc. ACM SIGIR Conf. Res., Develop. Inform. Retrieval*. Tampere, Finland: ACM, 2002, pp. 238–245.

[23] J. F. Canny, "Collaborative filtering with privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, 2002, pp. 45–57.

[24] W. Ahmad and A. Khokhar, "An architecture for privacy preserving collaborative filtering on web portals," in *Proc. IEEE Int. Symp. Inform. Assurance, Secur. (IAS)*. Washington, DC: IEEE Comput. Soc., 2007, pp. 273–278.

[25] J. Zhan, C. L. Hsieh, I. C. Wang, T. S. Hsu, C. J. Liau, and D. W. Wang, "Privacy-preserving collaborative recommender systems," *IEEE Trans. Syst. Man, Cybern.*, vol. 40, no. 4, pp. 472–476, Jul. 2010.

[26] M. Deng, "Privacy preserving content protection," Ph.D. dissertation, Katholieke Universiteit Leuven Faculty of Engineering, 2010.
[27] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–88, 1981.
[28] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Proxies for anonymous routing," in *Proc. Comput. Secur. Appl. Conf. (CSAC)*, San Diego, CA, Dec. 1996, pp. 9–13.
[29] D. Goldschlag, M. Reed, and P. Syverson, "Hiding routing information," in *Proc. Inform. Hiding Workshop (IH)*, 1996, pp. 137–150.
[30] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in *Proc. Conf. USENIX Secur. Symp.*, Berkeley, CA, 2004, pp. 21–21.
[31] B. N. Levine, M. K. Reiter, C. Wang, and M. Wright, "Timing attacks in low-latency mix systems," in *Proc. Int. Financial Cryptogr. Conf.* Springer-Verlag, 2004, pp. 251–265.
[32] K. Bauer, D. McCoy, D. Grunwald, T. Kohno, and D. Sicker, "Low-resource routing attacks against anonymous systems," University of Colorado, Tech. Rep., 2007.
[33] S. J. Murdoch and G. Danezis, "Low-cost traffic analysis of tor," in *Proc. IEEE Symp. Secur., Priv. (SP)*, 2005, pp. 183–195.
[34] B. Pfitzmann and A. Pfitzmann, "How to break the direct RSA implementation of mixes," in *Proc. Annual Int. Conf. Theory, Appl. of Cryptogr. Techniques (EUROCRYPT)*. Springer-Verlag, 1990, pp. 373–381.
[35] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proc. IEEE Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, 2010, pp. 1–21.
[36] M. Fredrikson and B. Livshits, "RePriv: Re-envisioning in-browser privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, May 2011, pp. 131–146.
[37] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "Coprivate query profile obfuscation by means of optimal query exchange between users," *IEEE Trans. Depend., Secure Comput.*, 2012. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TDSC.2012.16
[38] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.
[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[40] T. Ibaraki and N. Katoh, *Resource allocation problems: algorithmic approaches*. MIT Press, 1988.
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.
[42] "GroupLens research." [Online]. Available: http://www.grouplens.org
[43] "MovieLens 10M data set," Aug. 2011. [Online]. Available: http://www.grouplens.org/system/files/ml-10m-README.html